Sprachkontrolle im Spiegel der Maschinellen Übersetzung

Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung

Shaimaa Marzouk

Translation and Multilingual Natural Language Processing 20

language science press

Translation and Multilingual Natural Language Processing

In this series (see the complete series history at https://langsci-press.org/catalog/series/tmnlp):

9. Fox, Wendy. Can integrated titles improve the viewing experience? Investigating the impact of subtitling on the reception and enjoyment of film using eye tracking and questionnaire data.
10. Moran, Steven & Michael Cysouw. The Unicode cookbook for linguists: Managing writing systems using orthography profiles.
11. Fantinuoli, Claudio (ed.). Interpreting and technology.
12. Nitzke, Jean. Problem solving activities in post-editing and translation from scratch: A multi-method study.
13. Vandevoorde, Lore. Semantic differences in translation.
14. Bisiada, Mario (ed.). Empirical studies in translation and discourse.
15. Tra&Co Group (ed.). Translation, interpreting, cognition: The way out of the box.
16. Nitzke, Jean & Silvia Hansen-Schirra. A short guide to post-editing.
17. Hoberg, Felix. Informationsintegration in mehrsprachigen Textchats: Der Skype Translator im Sprachenpaar Katalanisch-Deutsch.
18. Kenny, Dorothy (ed.). Machine translation for everyone: Empowering users in the age of artificial intelligence.
19. Kajzer-Wietrzny, Marta, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.). Mediated discourse at the European Parliament: Empirical investigations.
20. Marzouk, Shaimaa. Sprachkontrolle im Spiegel der Maschinellen Übersetzung: Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung

ISSN: 2364-8899

Sprachkontrolle im Spiegel der Maschinellen Übersetzung

Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung

Shaimaa Marzouk

language science press

Shaimaa Marzouk. 2022. *Sprachkontrolle im Spiegel der Maschinellen Übersetzung: Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung* (Translation and Multilingual Natural Language Processing 20). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/371

ISBN: 978-3-96110-394-2 (Digital)
978-3-98554-052-5 (Hardcover)

ISSN: 2364-8899
DOI: 10.5281/zenodo.7031898
Source code available from www.github.com/langsci/371
Errata: paperhive.org/documents/remote?type=langsci&id=371

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
http://langsci-press.org

Storage and cataloguing done by FU Berlin

Inhaltsverzeichnis

Danksagung

Nach Jahren intensiver Arbeit liegt sie nun vor Ihnen: meine Dissertation. Damit ist es an der Zeit, mich bei denjenigen zu bedanken, die mich in dieser herausfordernden, aber auch immens bereichernden Phase begleitet und mir die Anfertigung dieser Promotionsschrift ermöglicht haben:

An erster Stelle – es mögen mir andere verzeihen – muss ich meine Doktormutter nennen. Frau Prof. Dr. Silvia Hansen-Schirra war zu jeder Zeit für mich da, motivierte mich unendlich und beseitigte alle aufkommenden Hindernisse. Dank ihres unerschöpflichen Fundus an thematischen und wissenschaftlichen Hinweisen lenkten mich unsere zahlreichen Gespräche stets in neue Sphären. Ohne ihre geduldige, inspirierende und konstruktive Unterstützung wäre diese Arbeit nicht gelungen. Ich bedanke mich herzlich bei der besten Doktormutter, die man sich wünschen kann.

Der gleiche Dank gilt meinem Betreuer, Herrn Prof. Dr. Christoph Rösener, für den kritischen Diskurs und die zielführenden Diskussionen. Dank seiner langjährigen Forschungs- und Praxiserfahrung im Bereich der Kontrollierten Sprache und CLCs eröffneten mir unsere Diskussionen andere Blickwinkel. In diesem Zusammenhang auch herzlichen Dank an ihn sowie an das Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung (IAI) der Universität des Saarlandes für die Forschungskooperation und die Zurverfügungstellung der Lizenz von CLAT.

Der gleiche Dank gilt meinem Betreuer, Herrn Prof. Dr. Bernd Meyer, für den intensiven und lehrreichen Austausch sowie das konstruktive Feedback, die mir halfen, Fragen aus der philosophischen Perspektive deutlicher zu sehen und zu klären. Ebenfalls bedanke ich mich vielmals bei ihm für die Gewährung großer wissenschaftlicher Freiräume für ein selbstständiges Arbeiten. Sowohl seine wissenschaftliche Betreuung als auch ständige Hilfsbereitschaft waren mir eine vielseitige Unterstützung. Herzlichen Dank.

In Zusammenhang mit der empirischen Arbeit unterstützen mich viele Menschen, ohne deren Hilfe eine solche Studie nicht durchführbar gewesen wäre:

vielen Dank an Frau Pruski und Herrn Behshad für die Bereitstellung unentbehrlicher Testdokumente und -quellen.

Ein besonders großes Dankeschön gilt Dr. Lorenz Kropfitsch, Bernd Schäfer, Sarah Signer, Patricia Graham und Kathleen Wallace für ihre herzliche und kompetente Unterstützung bei der Datenaufbereitung und den Testläufen.

Danken möchte ich außerdem den Probanden der empirischen Studie für ihre motivierte Teilnahme.

Für seine rege Aufmerksamkeit und Unterstützung bei der automatischen Evaluation bedanke ich mich herzlich bei Aaron L. F. Han von der Dublin City University.

Des Weiteren möchte ich mich bei Daniela Keller für ihre kompetente Unterstützung bei der statistischen Analyse, ihre endlose Geduld und die großartige positive Energie, die sie ausstrahlt, vielmals bedanken.

Großer Dank gebührt Ammar Suleiman für seine großartige Unterstützung bei den statistischen Tests, sein aufmerksames Zuhören und seine unglaublich freundliche und konstruktive Hilfe gerade in den stressigen Momenten.

Bei Agnieszka Surdyka möchte ich mich ferner für die mühevolle Arbeit des Korrekturlesens und vor allem für die reibungslose Kommunikation und die fröhliche Stimmung herzlich bedanken.

Darüber hinaus gilt mein Dank allen Verwandten, Studienkollegen und Freunden, die mich auch in schwierigen Zeiten unterstützt und immer wieder aufgeheitert haben. Dies war stets ein großer Rückhalt für mich, der als wichtiger Teil zum Erfolg meines Studiums beigetragen hat.

Tief verbunden und dankbar bin ich meiner Familie in Ägypten für ihre unglaubliche Geduld und ihr Verständnis bei der Anfertigung dieser Doktorarbeit.

Nicht minder aufreibend waren die vergangenen Jahre für meine Familie in Deutschland, die dieses Werk in allen Phasen mit jeder möglichen Unterstützung bedacht haben. Ohne Eure liebevolle Fürsorge wäre diese Arbeit nicht zu dem Werk geworden, welches sie heute ist. Für ihren unermüdlichen Beistand und ihre unendliche Geduld gilt mein voller Dank Thomas, Helga und Frank-Michael Wegener.

Ganz weit entfernt, doch mit einer ganz besonderen Bindung, sind zwei Personen, denen hier mein herauszustellender Dank gebührt; und zwar die zwei Personen, die mir den Durchhaltewillen und den Ehrgeiz für das Leben mitgaben: meine Großmutter und meine Mutter.

Zwar an letzter Stelle in dieser Aufzählung, dafür aber in meinem Herzen ganz weit vorn kommen meine tollen Kinder. Ich weiß, ihr habt in den letzten Jahren wegen mir auf viele Dinge verzichten müssen, und es tat mir oft sehr weh für euch. Liebe Salma, lieber Adam, ein riesiges Dankeschön aus tiefstem Herzen für euer Lächeln, das mich begleitet.

Dieses Buch widme ich Salma, Adam und Thomas.

Salma und Adam Marzouk-Wegener (2018): Die tollste Unterstützung beim Scannen, Bibliothek des FTSK, Germersheim.

Abkürzungsverzeichnis

AEMs	Automatische Evaluationsmetriken
CLC	Controlled-Language-Checker
CQ	Inhaltsqualität
ET	Eye-Tracking
GNMÜ	Google Neuronale Maschinelle Übersetzung
HMÜ	Hybride Maschinelle Übersetzung
KS	Kontrollierte Sprache
MÜ	Maschinelle Übersetzung
NMÜ	Neuronale Maschinelle Übersetzung
PE	Post-Editing
Q	Qualität
RBMÜ	Regelbasierte Maschinelle Übersetzung
SMÜ	Statistische Maschinelle Übersetzung
SQ	Stilqualität

1 Einleitung

> *The major obstacles to translating by computer are, as they have always been, not computational but linguistic.*
> Hutchins & Somers 1992: 2

1.1 Hintergrund und Motivation

Die Anwendung der Kontrollierten Sprache (KS) ist eine gängige Pre-Editing-Technik in der technischen Dokumentation. Eine KS wird wie folgt definiert: „an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style“ (Huijsen 1998: 2). Durch den Einsatz von lexikalischen, syntaktischen und stilistischen Restriktionen wird Ambiguität vermieden, die Satzstruktur vereinfacht und somit die Textkomplexität reduziert sowie Konsistenz und Standardisierung geschaffen. Das Ziel dabei ist es, die Textlesbarkeit, -verständlichkeit und -wiederverwendbarkeit zu erhöhen sowie die Qualität der menschlichen und maschinellen Übersetzung (MÜ) zu verbessern und die Übersetzungskosten zu reduzieren (vgl. Hutchins & Somers 1992: 4; Lehrndorfer 1996b; Huijsen 1998; Nyberg u. a. 2003: 245; Drewer & Ziegler 2014: 197). Im Fokus der vorliegenden Studie liegt die Beziehung zwischen der KS und der maschinellen Übersetzbarkeit. Die MÜ gilt als „one of the most interesting computational applications of Controlled Language“ (Nyberg u. a. 2003: 254f.). Die Grundidee des KS-Einsatzes im Kontext der MÜ ist es, dass je einfacher der Text auf linguistischer Ebene ist, desto höher die Qualität seiner MÜ ausfällt, wie Hutchins und Somers es bereits 1992 auf den Punkt brachten: „The major obstacles to translating by computer are, as they have always been, not computational but linguistic.“ (Hutchins & Somers 1992: 2) Auch Regeln, die die Verständlichkeit (und nicht die Übersetzbarkeit) im Fokus haben, optimieren die Verständlichkeit hauptsächlich auf dreierlei Weise, nämlich durch die Vereinfachung der Satzstruktur, Reduzierung der Satzkomplexität bzw. Vermeidung der Ambiguität; dadurch tragen sie zudem indirekt zur Verbesserung der Übersetzbarkeit bei (vgl. Fiederer & O’Brien 2009: 53).

Bereits Mitte der 80er Jahre entwickelte Caterpillar CTE „Caterpillar Technical English“. Im Mittelpunkt standen die Optimierung der Text- und Übersetzungsqualität sowie die Reduzierung der Übersetzungskosten. In der Tat gelang es Caterpillar mithilfe der CTE in Verbindung mit einem regelbasierten MÜ-System (RBMÜ), ein großes Textvolumen zu verfassen und es in mehr als 30 Sprachen zu übersetzen. (Kamprath u. a. 1998; Drewer & Ziegler 2014: 212) Ebenfalls ermöglichte 1979 der Einsatz der KS (Multinational Customised English, MCE) in Kombination mit einem RBMÜ-System Xerox, das in mehr als 36 Ländern tätig war, seine Übersetzungsproduktivität zu vervierfachen (Elliston 1979). Angesichts dieses Erfolgs begannen weitere Organisationen firmen- und branchenspezifische KS zu entwickeln, wie z. B. das „NCR Fundamental English“, das „Plain English Program“ (PEP) und die „International Language for Servicing and Maintenance“ (ILSAM) (vgl. Schwanke 1991: 42). Zudem fand die KS Anklang bei multinationalen Konzernen, wie Kodak (Kodak International Service Language), Nortel (Nortel Standard English), Sun (Sun Controlled English), Ericsson (Ericsson English) und KANT (KANT Controlled English) (Drewer & Ziegler 2014: 211).

In der Forschung zeigten mehrere Studien über das letzte Vierteljahrhundert, dass sich die Anwendung der KS aus verschiedenen Perspektiven positiv auf die MÜ auswirkt. Kamprath u. a. (1998) belegten einen signifikanten positiven Einfluss von CTE auf die MÜ-Produktivität. Reuther (2003) kam zum Ergebnis, dass die Implementierung von KS-Regeln die Lesbarkeit und Übersetzbarkeit der maschinell übersetzten Texte verbessern könne. Eine weitere Studie ergab, dass die Kontrolle des Ausgangstexts einen großen Einfluss auf die Genauigkeit des MÜ-Outputs eines wissensbasierten interlingualen MÜ-Systems hat (Nyberg & Mitamura 1996). Bernth & Gdaniec (2001) führten 26 Regeln für Englisch als Ausgangssprache ein, die unterschiedliche Texteigenschaften abdecken, mit dem Ziel, die maschinelle Übersetzbarkeit zu verbessern. Sie testeten diese Regeln mit verschiedenen kommerziellen MÜ-Systemen und gaben an, dass sie auf verschiedene MÜ-Systeme und Sprachenpaare generalisierbar seien (ebd.). Darüber hinaus untersuchten weitere Studien die Auswirkungen der KS auf das Post-Editing. O’Brien (2006) stellte fest, dass die KS die Post-Editing-Zeit verkürzt. Ein weiterer positiver Effekt war in Zusammenhang mit der Post-Editing-Produktivität nachweisbar (Aikawa u. a. 2007).

Die meisten KS-Studien haben den Einfluss vollständiger KS auf die (maschinelle) Übersetzbarkeit ganzheitlich untersucht (z. B. Spyridakis u. a. 1997; Kamprath u. a. 1998; Bernth 1999; Nyberg u. a. 2003: 254f.). Die Ergebnisse dieser Untersuchungen liefern ein Gesamtbild des KS-Effekts (inkl. Regeln zur Verbesserung der Textlesbarkeit, -verständlichkeit und -übersetzbarkeit), bei dem der positive Effekt einiger Regeln den negativen Effekt anderer Regeln überschatten

kann, was zu einem verzerrten Endergebnis führt und den individuellen Einfluss der einzelnen Regeln nicht erkennen lässt. Nur eine begrenzte Anzahl von Studien befasste sich mit der Analyse des Einflusses einzelner KS-Regeln. Die Ergebnisse dieser Studien (O'Brien 2006; Roturier 2006; Roturier u. a. 2012) zeigen, dass die KS-Regeln die MÜ auf verschiedene Weise und in unterschiedlichem Maße beeinflussten. Diese Studien untersuchten KS-Regeln der englischen Sprache. Roturier u. a. (2012) analysierten die Auswirkungen der KS-Regeln auf die MÜ-Qualität mithilfe von automatischen Evaluationsmetriken. Die Experimente wurden mit einem phrasenbasierten System (PBMÜ) für die Zielsprachen Französisch und Deutsch durchgeführt. In einer weiteren Studie konzentrierte sich Roturier (2006) ebenfalls auf die gleichen Zielsprachen, verwendete jedoch ein RBMÜ-System mit dem Ziel, die Auswirkungen von KS-Regeln auf die Verständlichkeit zu analysieren. O'Brien (2006) untersuchte die Auswirkungen von KS-Regeln auf den Post-Editing-Aufwand für Deutsch als Zielsprache bei einem anderen RBMÜ-System.

Die Untersuchung der individuellen Auswirkungen der KS-Regeln bei verschiedenen Systemen ermöglicht einen effizienten Einsatz der sich positiv auswirkenden Regeln. Ferner – je nach Implementierungskontext – besteht die Problematik: „some writing rules may even do more harm than good" (Nyberg u. a. 2003: 257). Dementsprechend können anhand solcher Untersuchungen sich negativ auswirkende Regeln ausgeschlossen werden. Darüber hinaus hat eine Reduktion der Regeln auf das Wesentliche den Vorteil, dass potenzielle Schwächen der KS-Anwendung begrenzt werden: Der Einsatz umfangreicher KS-Regeln kann mit einem Regel-Usability-Problem verbunden sein, die Autorenproduktivität beeinträchtigen (Mitamura 1999), den Schreibprozess übermäßig komplex gestalten bzw. übermäßig in diesen eingreifen (O'Brien & Roturier 2007) und zur Ablehnung von den Autoren führen (Doherty 2012: 31). Eine weitere Nebenwirkung der KS besteht darin, dass die Textakzeptabilität durch den Einsatz der Regeln beeinträchtigt werden kann (Roturier 2006). Der Stil der KS kann von manchen Rezipienten als unästhetisch empfunden werden (Lehrndorfer & Reuther 2008: 112f.). Im Bereich der Leichten Sprache (LS), als einer Varietät der KS (siehe §2.2.1), diskutierten Hansen-Schirra & Maaß (2020) den Trade-off zwischen der Verständlichkeit und der Akzeptanz, der oft in dieser Sprachvarietät beobachtet wird. Sie betonten die Bedeutung des Akzeptanzfaktors für die Zielgruppen dieser Sprachvarietät und modellierten folglich die LS „LS+", die sowohl verständlich als auch akzeptabel ist (ebd.). Außerdem kann die Anwendung von großen KS-Regelsätzen – trotz der Verwendung von einem KS-Checker – aus Zeit- und Ressourcengründen schwierig ausfallen (Govyaerts 1996; O'Brien & Roturier 2007). Schließlich wird die technische Dokumentation nicht selten von den jeweiligen

Fachabteilungen bzw. Fachexperten verfasst, die über begrenztes linguistisches Wissen zum Verstehen und der Umsetzung aller Regeln verfügen (Van der Eijk u. a. 1996; Aranberri & Roturier 2009). All diese potenziellen Schwächen unterstreichen die Notwendigkeit einer sorgfältigen Identifizierung der für bestimmte Ziele tatsächlich erforderlichen Regeln.

Eine empirische Analyse der Auswirkungen der einzelnen Regeln ist demzufolge erforderlich, um über die Anwendung bzw. Nicht-Anwendung der Regeln auf solider Basis entscheiden zu können. Allerdings stellt die Ermittlung der Auswirkung einer bestimmten Regel innerhalb eines Regelsatzes eine Herausforderung dar, die von Douglas & Hurst (1996: 2) folgendermaßen beschrieben wurde:

> While it might be possible to evaluate the quality of a document conforming to a specific CL [...], this does not allow us to say anything about the effects of particular elements in the definition of the CL, some of which may impose burdens on the writer disproportionate to their benefits to the reader. (Douglas & Hurst 1996: 2)

In der Tat ist eine solche empirische Studie mit mehreren Herausforderungen verbunden, was die begrenzte Anzahl an Studien in diesem Bereich erklärt. Die Hauptherausforderungen umfassen:

- die Entwicklung einer Technik, die eine Isolierung der Auswirkung der einzelnen Regeln ermöglicht;
- die Aufbereitung der Daten nach klar definierten Kriterien, um linguistische Schwierigkeiten, die für das getestete Problem irrelevant sind (sog. Noise), zu eliminieren bzw. möglichst zu reduzieren (vgl. King & Falkedal 1990);
- die Umsetzung der Regeln nach einem begründeten festgelegten Muster, da die Regelumsetzungsmuster die Ergebnisse erheblich beeinflussen (Roturier 2006: 74);
- die Arbeit mit einem Datensatz, mit dem diese Forschungsherausforderungen bewältigt werden können und der gleichzeitig über eine für die statistische Auswertung ausreichende Größe verfügt.

Im Allgemeinen sind MÜ-Studien im Bereich der KS aufgrund der Anzahl der unabhängigen Variablen ein komplexes Forschungsgebiet. Da die MÜ-Qualität

je nach Sprachenpaar, Übersetzungsrichtung, Domäne und angewendetem MÜ-System variiert, ist es zu erwarten, dass die Auswirkung jeder KS-Regel auf den MÜ-Output zusammen mit diesen Variablen ebenfalls variiert.

In den letzten Jahren wurde der neuronale MÜ-Ansatz (NMÜ) eingeführt; ein Ansatz, der mithilfe der künstlichen Intelligenz auf Basis neuronaler Netzmodelle versucht, das menschliche Gehirn bzw. Denken nachzuahmen, wodurch er sich signifikant von früheren Ansätzen abhebt. Das Vorhaben dieser Studie wurde durch die Entwicklung des NMÜ-Ansatzes in Gang gebracht. Angesichts des Quantensprungs in der MÜ-Qualität und der damit verbundenen Veränderung der Art der aufgetretenen Fehler ist es an der Zeit, die Auswirkungen der KS auf die maschinelle Übersetzbarkeit bei der NMÜ zu überprüfen. Mehrere Studien belegen den Erfolg dieses Ansatzes bei der Erstellung hochqualitativer Übersetzungen im Vergleich zu den früheren Ansätzen sowohl für das hier untersuchte Sprachenpaar als auch für weitere Sprachenpaare (vgl. Bentivogli u. a. 2016; Wu u. a. 2016; Castilho, Moorkens, Gaspari, Sennrich u. a. 2017; Toral & Sanchez-Cartagena 2017; Popović 2018). Während die vorherigen Ansätze mit schwerwiegenden Fehlern (u. a. im Bereich der Morphologie und Grammatik, wie z. B. Wortstellungsfehlern) zu kämpfen haben, kann die NMÜ diese Schwierigkeiten lösen und darüber hinaus eine im Wesentlichen flüssige Übersetzung liefern (vgl. Bentivogli u. a. 2016; Toral & Sanchez-Cartagena 2017; Van Brussel u. a. 2018). Eine relative Schwäche zeigt die NMÜ dennoch bei den Qualitätskriterien Genauigkeit bzw. Adäquatheit, inkl. wenig transparenten Fehlern wie z. B. Fehlübersetzungen, Auslassungen und stilistischen Fehlern, die mit einem hohen Post-Editing-Aufwand verbunden sind (Van Brussel u. a. 2018; Volk 2018; Vardaro u. a. 2019). Solche Qualitätskriterien bedürfen einer Humanevaluation, in der die Evaluatoren trotz der hohen Flüssigkeit der Übersetzung (feine) Genauigkeits-, Adäquatheits- bzw. stilistische Fehler identifizieren, denn AEMs sind nicht immer in der Lage, sie differenziert auszuwerten (vgl. Müller u. a. 2018; Shterionov u. a. 2018). Vor diesem Hintergrund dürfen die Genauigkeit bzw. Adäquatheit und der Stil bei der MÜ-Evaluation nicht außer Acht gelassen werden, und das, obwohl der Stil kein Hauptziel der KS ist. Die vorliegende Studie beleuchtet die einzelnen Regeln eingehender, um ihre Wirkung auf die MÜ-Qualität sowohl stilistisch als auch inhaltlich im Sinne der Verständlichkeit und Genauigkeit empirisch zu testen. Auf diese Weise kann die Entscheidung über die Anwendung einer Regel kontextabhängig unter Betrachtung aller Qualitätsperspektiven auf solider empirischer Basis getroffen werden.

Wie diese kurze Darstellung der MÜ-Studien im Bereich der KS zeigt, lag der Forschungsfokus auf der englischen KS im Kontext der RBMÜ, SMÜ bzw. PBMÜ. Das kontrollierte Deutsch wurde selten erforscht und – nach bestem Wissen –

bislang nicht im Kontext der NMÜ betrachtet. Die vorliegende Arbeit versucht, durch die Abdeckung des NMÜ-Ansatzes und den Vergleich seines KS-Einflusses mit denen früherer Ansätze, diese Forschungslücke zu schließen. Wie oben erwähnt, wird die MÜ-Forschung im Bereich der KS im Allgemeinen als komplex angesehen, da die MÜ-Qualität sich je nach Sprachenpaar, Übersetzungsrichtung, Domäne und angewendetem MÜ-System unterscheidet und davon abhängig die Auswirkung jeder KS-Regel auf den MÜ-Output zusammen mit diesen Variablen variiert. Dieser Komplexität wurde begegnet, indem die Variablen Sprachenpaar, Übersetzungsrichtung und Domäne konstant gehalten wurden, mit dem Ziel, die individuellen Auswirkungen von neun KS-Regeln auf den Output von fünf MÜ-Systemen der RBMÜ-, SMÜ-, HMÜ- und NMÜ-Ansätze auf verschiedenen Ebenen zu analysieren und gegenüberzustellen. Die Studie wurde im technischen Bereich bzw. für die technische Dokumentation durchgeführt, da dies das Hauptanwendungsgebiet der KS darstellt. Aufgrund des Mangels an empirischen Untersuchungen der KS-Regeln der deutschen Sprache befasst sich die Arbeit mit dem Sprachenpaar Deutsch>Englisch. Angesichts der demonstrierten Notwendigkeit der KS-Untersuchung auf Regelebene, die gleichzeitig mit einer Reihe von Herausforderungen gekoppelt ist, widmet sich die Studie dieser Untersuchungsebene. Es wurde dafür ein Mixed-Methods-Triangulationsansatz angewendet und versucht, durch die Details der Analyseverfahren, den unterschiedlichen Herausforderungen gerecht zu werden.[1]

1.2 Ziel der Studie

Angesichts der jüngsten MÜ-Entwicklungen besteht das Ziel der Studie darin, den Einfluss *einzelner* KS-Regeln der deutschen Sprache auf die Qualität des MÜ-Outputs in der englischen Sprache bei verschiedenen MÜ-Ansätzen (RBMÜ, SMÜ, HMÜ und NMÜ) empirisch zu untersuchen und zu vergleichen. Der Einfluss der KS-Regeln auf die MÜ-Outputqualität wurde in Bezug auf die Anzahl und Typen der aufgetretenen Fehler, Stil- und Inhaltsqualität sowie AEM-Scores (Automatic Evaluation Metric) nach einem Mixed-Methods-Triangulationsansatz erforscht. Die Anzahl und Typen der aufgetretenen Fehler wurden im Rahmen einer Fehlerannotation basierend auf der Fehlertaxonomie von Vilar u. a. (2006) ermittelt (siehe §4.5.4). Die Stil- und Inhaltsqualität wurden in Anlehnung an Hutchins & Somers (1992: 163) definiert, wobei die Inhaltsqualität die Attribute Genauigkeit und Verständlichkeit umfasst und die Stilqualität die Idiomatik

[1] Alle diesbezüglichen Details sind unter §3.5.3 „Forschungsherausforderungen der KS-Untersuchungen auf Regelebene“ ausführlicher diskutiert.

der MÜ, die Eignung der MÜ für die Intention des Inhalts sowie die korrekte bzw. klare orthografische Darstellung der MÜ abdeckt (siehe Definitionen unter §4.5.5.1).[2] Die Analyse der Stil- und Inhaltsqualität erfolgte anhand einer Humanevaluation. Die Scores von TERbase (Snover u. a. 2006) und hLEPOR (Han u. a. 2013) wurden in einer automatischen Evaluation gemessen (siehe §4.5.6). Um das Ziel der Studie zu realisieren, wurden die zwei Szenarien „vor Anwendung der KS-Regel" (nachstehend „vor-KS" genannt) und „nach Anwendung der KS-Regel" (nachstehend „nach-KS" genannt) auf Satzebene verglichen. Der Vergleich fand auf vier Ebenen statt:

(1) *Auf Sprachenpaarebene*: ein Vergleich der Szenarien vor-KS vs. nach-KS im gesamten Datensatz.

(2) *Auf Regelebene*: der Datensatz wurde nach Regel aufgeteilt (d. h. bei jeder Regel sind alle MÜ-Systeme enthalten). Die Szenarien vor-KS vs. nach-KS wurden bei den einzelnen Regeln verglichen.

(3) *Auf MÜ-Systemebene*: der Datensatz wurde nach MÜ-System aufgeteilt (d. h. bei jedem System sind alle Regeln enthalten). Die Szenarien vor-KS vs. nach-KS wurden bei den einzelnen MÜ-Systemen verglichen.

(4) *Auf Regel- und MÜ-Systemebene*: der Datensatz wurde nach Regel aufgeteilt und anschließend innerhalb jeder Regel nach MÜ-System aufgeteilt. Die Szenarien vor-KS vs. nach-KS wurden bei den einzelnen MÜ-Systemen innerhalb jeder Regel verglichen.

1.3 Forschungsfragen und Hypothesen

Die Hauptfragestellung der Studie lautet: Wo liegen die Unterschiede im maschinellen Übersetzungsoutput vor vs. nach der Anwendung der einzelnen analysierten KS-Regeln in Bezug auf die aufgetretenen Fehler, Stil- und Inhaltsqualität

[2] Die Begriffe Verständlichkeit, Intention und Stil werden in verschiedenen Disziplinen und Subdisziplinen behandelt. Für jeden dieser Begriffe existieren dementsprechend zahlreiche konkurrierende Definitionen, die aus Platzgründen nicht alle präsentiert und gegeneinander abgewogen werden können. Im Rahmen dieser Studie wurden diese Begriffe aus computerlinguistischer Perspektive betrachtet und nach den Definitionen von Hutchins & Somers (1992), die seit den neunziger Jahren in der MÜ-Evaluation häufig herangezogen werden, bei der Evaluation umgesetzt. Unter §4.5.5.1 werden die Definitionen genauer erläutert und ihre Auswahl begründet. Eine weiterführende Diskussion dieser Begriffe würde den Rahmen dieser Arbeit sprengen.

sowie AEM-Scores regel- und systemübergreifend, regelspezifisch sowie systemspezifisch?

Um diese Hauptfragestellung umfassend zu beantworten, musste sie in detaillierte Fragestellungen unterteilt werden, die neun Analysefaktoren zugeordnet wurden. Im Folgenden sind die Analysefaktoren zusammen mit den Fragestellungen, Hypothesen, Analyseebenen sowie angewandten Analysemethoden in einer Übersicht aufgeführt:

Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach dem Einsatz der KS-Regel

Anhand des ersten Analysefaktors wird die Fehleranzahl auf Basis der Fehlerannotation vor vs. nach der KS-Anwendung innerhalb der KS-Stelle[3] verglichen (Fragestellung [1]):

Fragestellung Gibt es einen Unterschied in der Fehleranzahl nach dem Einsatz der KS-Regel im Vergleich zu vor dem Einsatz?

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene
- Regel- und MÜ-Systemebene

Analysemethode Fehlerannotation

Hypothesen

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor- und nach-KS.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor- und nach-KS.

Zweiter Analysefaktor: Vergleich der Fehleranzahl vor vs. nach dem Einsatz der KS-Regel außerhalb der KS-Stelle bei einer korrekten Übersetzung der KS-Stelle

Bei einer korrekten Übersetzung der KS-Stelle sowohl vor als auch nach der KS-Anwendung wird anhand des zweiten Analysefaktors die Fehleranzahl außerhalb der KS-Stelle vor vs. nach der KS-Anwendung verglichen. Ziel hierbei ist es,

[3] Die KS-Stelle ist der Teil des Ausgangssatzes, der bei dem Einsatz der KS-Regel modifiziert werden muss, und seine Äquivalenz im Zielsatz. Mehr dazu unter Abschnitt §4.5.2.1.

eine weitere potenzielle Wirkung der Regeln außerhalb der KS-Stelle – trotz der korrekt übersetzten KS-Stelle – einzubeziehen (Fragestellung [2]):

Fragestellung Stieg die Fehleranzahl außerhalb der KS-Stelle nach dem Einsatz der KS im Vergleich zu davor, obwohl die MÜ innerhalb der KS-Stelle sowohl vor als auch nach dem Einsatz der KS richtig waren?

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene

Analysemethode Fehlerannotation

Dritter Analysefaktor: Aufteilung der Annotationsgruppen

In der Studie werden die Ergebnisse der Fehlerannotation auf Basis der Existenz und Nicht-Existenz von Fehlern in vier Gruppen unterteilt, bezeichnet als „Annotationsgruppen". Diese Annotationsgruppen sind: (1) RR – MÜ ist vor und nach der Anwendung der KS-Regel fehlerfrei; (2) FF – MÜ beinhaltet vor und nach der Anwendung der KS-Regel Fehler; (3) RF – MÜ ist nur vor der Anwendung der KS-Regel fehlerfrei; (4) FR – MÜ ist nur nach der Anwendung der KS-Regel fehlerfrei. Bei dem dritten Analysefaktor geht es um die prozentuale Aufteilung dieser Annotationsgruppen (Fragestellung [3]):

Fragestellung Wie hoch ist der Prozentsatz jeder Annotationsgruppe?

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene
- Regel- und MÜ-Systemebene

Analysemethode Fehlerannotation

Vierter Analysefaktor: Vergleich der Fehlertypen vor vs. nach dem Einsatz der KS-Regel

Durch den vierten Analysefaktor werden nach einer Fehlertaxonomie 13 Fehlertypen vor vs. nach der KS-Anwendung verglichen, um herauszufinden, ob bestimmte Fehlertypen in Verbindung mit der Regelanwendung ab- oder zunahmen (Fragestellung [4]):

Fragestellung Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach dem Einsatz der KS-Regel?

Davon wird abgeleitet,

(1) ob bestimmte Fehlertypen, die vor dem Einsatz der KS-Regel existierten, nach dem Einsatz der KS-Regel eliminiert bzw. reduziert wurden;

(2) ob bestimmte Fehlertypen erst nach dem Einsatz der KS-Regel erschienen bzw. deutlich stiegen (im Vergleich zu vor dem Einsatz der KS-Regel).

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene
- Regel- und MÜ-Systemebene

Analysemethode Fehlerannotation

Hypothesen

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach dem Einsatz der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach dem Einsatz der KS-Regel.

Fünfter Analysefaktor: Vergleich der Qualität vor vs. nach dem Einsatz der KS-Regel

Bei dem fünften Analysefaktor werden die im Rahmen der Humanevaluation vergebenen Qualitätscores vor vs. nach der KS-Anwendung verglichen (Fragestellung [5]):

Fragestellung Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach dem Einsatz der KS-Regel im Vergleich zu vor dem Einsatz?

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene
- Regel- und MÜ-Systemebene

Analysemethode Humanevaluation

Hypothesen

H0 – Es gibt keinen Unterschied in der Qualität vor vs. nach dem Einsatz der KS-Regel.

H1 – Es gibt einen Unterschied in der Qualität vor vs. nach dem Einsatz der KS-Regel.

Sechster Analysefaktor: Vergleich der Qualität vor vs. nach dem Einsatz der KS-Regel auf Annotationsgruppenebene

Auf Basis einer Triangulation der Ergebnisse der Fehlerannotation mit denen der Humanevaluation wird bei den einzelnen Annotationsgruppen untersucht, ob die Stil- und Inhaltsqualität vor vs. nach der KS-Anwendung sanken bzw. stiegen. Während ein Qualitätsanstieg bei der Gruppe FR bzw. ein Qualitätsrückgang bei der Gruppe RF erwartet wird, ist es von Interesse zu analysieren, wie die Qualitätsveränderung bei zwei fehlerfreien MÜ (Gruppe RR) bzw. bei zwei fehlerhaften MÜ (Gruppe FF) ausfiel (Fragestellung [6]):

Fragestellung Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach dem Einsatz der KS-Regel im Vergleich zu vor dem Einsatz?

Davon wird abgeleitet,

(1) ob bei der Gruppe RR die Stil- bzw. Inhaltsqualität vor bzw. nach dem Einsatz der KS-Regel höher ist, obwohl die MÜ in beiden Fällen fehlerfrei ist;

(2) ob bei der Gruppe FF die Stil- bzw. Inhaltsqualität vor bzw. nach dem Einsatz der KS-Regel höher ist, obwohl die MÜ in beiden Fällen Fehler beinhaltet;

(3) ob bei der Gruppe RF die Stil- bzw. Inhaltsqualität nach dem Einsatz der KS-Regel stieg, obwohl die MÜ nach dem Einsatz der KS-Regel Fehler beinhaltet und davor fehlerfrei war;

(4) ob bei der Gruppe FR die Stil- bzw. Inhaltsqualität nach dem Einsatz der KS-Regel sank, obwohl die MÜ nach dem Einsatz der KS-Regel fehlerfrei ist und davor Fehler beinhaltete.

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene

Analysemethode Fehlerannotation, Humanevaluation

Hypothesen

H0 – Bei den Annotationsgruppen gibt es keinen Unterschied in der Qualität vor vs. nach dem Einsatz der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Unterschied in der Qualität vor vs. nach dem Einsatz der KS-Regel.

Siebter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Mithilfe der triangulierten Ergebnisse der Fehlerannotation und der Humanevaluation wird beim siebten Analysefaktor der potenzielle Zusammenhang zwischen den Fehlertypen und der Qualität anhand einer Korrelationsberechnung untersucht (Fragestellung [7]):

Fragestellung Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS − Qualität vor KS)?

Analyseebene

- Sprachenpaarebene

- Regelebene
- MÜ-Systemebene
- Regel- und MÜ-Systemebene

Analysemethode Fehlerannotation, Humanevaluation

Hypothesen

H0 – Es gibt keinen Unterschied in der Qualität vor vs. nach dem Einsatz der KS-Regel.

H1 – Es gibt einen Unterschied in der Qualität vor vs. nach dem Einsatz der KS-Regel.

Achter Analysefaktor: Vergleich der Scores der AEMs vor vs. nach dem Einsatz der KS-Regel

Die bei der Humanevaluation angegebenen Referenzübersetzungen werden bei der automatischen Evaluation zur Berechnung der AEM-Scores angewendet. Anhand des achten Analysefaktors werden die Scores von TERbase und hLEPOR vor vs. nach der KS-Anwendung verglichen (Fragestellung [8]):

Fragestellung Gibt es einen Unterschied in den Scores von TERbase bzw. hLEPOR nach dem Einsatz der KS-Regel im Vergleich zu vor dem Einsatz?

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene

Analysemethode Humanevaluation, Automatische Evaluation

Hypothesen

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach dem Einsatz der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach dem Einsatz der KS-Regel.

Neunter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualitätsdifferenz

Auf Basis einer Triangulation der Ergebnisse der Humanevaluation mit denen der automatischen Evaluation wird beim neunten Analysefaktor der potenzielle Zusammenhang zwischen den Differenzen der AEM-Scores und der Qualitätsdifferenz anhand einer Korrelationsberechnung analysiert (Fragestellung [9]):

Fragestellung Besteht ein Zusammenhang zwischen der Differenz der Scores von TERbase bzw. hLEPOR (Mittelwert der Scores nach KS − Mittelwert der Scores vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS − Qualität vor KS)?

Analyseebene

- Sprachenpaarebene
- Regelebene
- MÜ-Systemebene

Analysemethode Humanevaluation, Automatische Evaluation

Hypothesen

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

1.4 Struktur der Studie

Nachdem der Hintergrund der Arbeit, die Motivation zur Untersuchung, das Ziel sowie die Fragestellungen und Hypothesen der Studie in diesem Kapitel demonstriert wurden, folgen die drei Hauptteile der Arbeit:

Der erste Hauptteil beschäftigt sich mit der theoretischen Grundlage der Kontrollierten Sprache (Kapitel 2) und der Maschinellen Übersetzung (Kapitel 3). In Kapitel 2 wird auf die Definition, die Ziele, den Aufbau und die Entwicklung der KS näher eingegangen. Anschließend erfolgt eine detaillierte Betrachtung des Kontrollierten Deutsch bzw. der tekom-Leitlinie als Basis für das Untersuchungsobjekt der Arbeit. Daraufhin werden die Stärken und Schwächen der KS diskutiert. Zum Schluss werden die CL-Checker und ihre Funktionsweise beschrieben. Kapitel 3 beginnt mit einer Erläuterung des MÜ-Begriffs und dem

Motiv eines KS-Einsatzes im Kontext der MÜ. Anschließend wird die Entwicklung der MÜ und ihrer Ansätze dargestellt. Darauffolgend befasst sich das Kapitel mit der Thematik der MÜ-Qualität sowie den verschiedenen Methoden der MÜ-Qualitätsevaluation und dem Evaluationsdesign. Zum Schluss werden MÜ-Studien im Kontext der KS diskutiert.

Der zweite Hauptteil widmet sich der angewandten Methodologie (Kapitel 4). Zunächst wird die Forschungsmethodik gefolgt von der Operationalisierung und der Validität der Arbeit genauer erläutert. Im Abschnitt Studiendesign wird die Auswahl des analysierten Sprachenpaars, der untersuchten Regeln und MÜ-Systeme sowie der Aufbau des Datensatzes ausführlich dargestellt. Zudem werden die methodischen Überlegungen und Testläufe, die zum Design der drei implementierten Evaluationsmethoden geführt haben, sowie die Vorgehensweisen und schließlich die Struktur der Ergebnisse der durchgeführten Analysen detailliert präsentiert.

Der letzte Hauptteil der Arbeit befasst sich mit den Studienergebnissen und umfasst Kapitel 5 und 6. In Kapitel 5 werden die Ergebnisse auf vier Analyseebenen in drei Unterkapiteln darstellt: Im ersten Unterkapitel werden zunächst der Datensatz zusammen mit den Ergebnissen einer allgemeinen Analyse der Entwicklung der MÜ-Qualität, des Interrater- und Intrarater-Agreements sowie der Profildaten der Teilnehmer präsentiert. Zudem ist eine Darstellung der Ergebnisse auf Sprachenpaarebene in diesem Unterkapitel vorgesehen. Im zweiten und dritten Unterkapitel werden die Ergebnisse auf Systemebene bzw. auf Regelebene sowie auf Regel- und Systemebene demonstriert. Abschließend liefert Kapitel 6 eine Zusammenfassung sowie eine Diskussion der Studienergebnisse.

Abschließend folgt das Fazit in Kapitel 7, in dem die Schlussfolgerungen dargestellt und die Implikationen rück- und ausblickend demonstriert werden, zudem erfolgt eine Präsentation des Beitrags, der Einschränkungen der Studie sowie von Ideen zur Anregung von zukünftiger Forschung; durch das Schlusswort wird die Arbeit abgerundet. Das Literaturverzeichnis und die Anhänge machen die verbleibenden Elemente aus.

2 Kontrollierte Sprache

Unternehmen, die mehrsprachige Dokumentation erstellen, […] unterscheiden sich heute nicht mehr darin, ob sie eine Kontrollierte Sprache benutzen oder nicht, sondern nur noch darin, wie umfassend die Sammlung der Formulierungsregeln ist und wie streng sie eingehalten werden. (Göpferich 2007a: 19)

2.1 Einleitung

Das vorliegende Kapitel befasst sich mit dem Thema Kontrollierte Sprache. Nach einer Erläuterung des Begriffs der Kontrollierten Sprache (KS), wird näher auf die Definition, die Ziele und den Aufbau der KS eingegangen. Anschließend werden die Entstehung und Entwicklung der KS beleuchtet sowie Beispiele von bekannten Kontrollierten Sprachen in verschiedenen Sprachen genannt. Da ausgewählte Regeln des Kontrollierten Deutsch bzw. der tekom-Leitlinie das Untersuchungsobjekt darstellen, werden das Kontrollierte Deutsch und die tekom-Leitlinie im Anschluss detaillierter betrachtet. Ferner werden die Stärken und Schwächen der KS gegenübergestellt. Da der Einsatz von einer KS in den Unternehmen in der Regel durch eine Software (CL-Checker) unterstützt wird, wird diese Software bzw. ihre Funktionsweise anschließend beschrieben.

2.2 Kontrollierte Sprache: Eine Einführung

2.2.1 Kontrollierte Sprache – Begriffsbestimmung und Definition

Die „Kontrollierte Sprache“ ist eine wörtliche Übersetzung der englischen Bezeichnung „Controlled Language“. Eine Sprache kann nicht kontrolliert werden.

In der Fachkommunikation ist „kontrollieren“ keine treffende Übersetzung für das englische Verb „control“. Die Bezeichnungen „regulierte“, „gesteuerte“ oder „standarisierte Sprache“ wären zutreffendere Übersetzungen (vgl. Schubert 1999: 434; Göpferich 2007a). Ein weiterer Aspekt des Begriffs ist, dass das Wort „Kontrollieren“ im Deutschen negativ konnotiert ist. Dennoch blieb die Bezeichnung „Kontrollierte Sprache“ sowohl in der Forschung als auch in der Industrie die meist verwendete Bezeichnung. Aus diesem Grund wird die Bezeichnung „Kontrollierte Sprache“ in der vorliegenden Arbeit übernommen.

Grundsätzlich kann jede festgelegte Regelsammlung bzw. jeder Redaktionsleitfaden als eine Kontrollierte Sprache von unterschiedlichen Umfängen betrachtet werden (vgl. Göpferich 2007a). Anhand einer genauen Betrachtung der Definitionen der Kontrollierten Sprache können die konkreten Merkmale einer Kontrollierten Sprache (KS) abgeleitet werden:

Nach Göpferich (2008: 366) sind Kontrollierte Sprachen „Subsysteme natürlicher Sprachen, deren Wortschatz und zulässige grammatische Konstruktionen eine Teilmenge des Wortschatzes bzw. der möglichen grammatischen Konstruktionen der unkontrollierten natürlichen Sprache darstellen, aus denen sie abgeleitet sind“. Eine ähnliche Definition liefern Drewer & Ziegler (2014: 192), in der sie unter einer KS „ein Teilsystem einer natürlichen Sprache mit den folgenden Bestandteilen: Regelwerk mit festen Schreibregeln (v. a. Vorgaben zur Formulierung von Sätzen) sowie vorgegebener Wortschatz aus Basiswortschatz und Fachwortschatz in Form eines Lexikons“ verstehen. Auf Basis der beiden Definitionen zeichnet sich eine Kontrollierte Sprache hauptsächlich dadurch aus, dass sie einer bestimmten natürlichen Sprache entstammt, wobei der Unterschied zwischen der KS und ihrer natürlichen Sprache in dem eingeschränkten Lexikon und der vereinfachten Syntax liegt. Eine Kontrollierte Sprache besteht konkret aus den folgenden Bestandteilen: einem spezifischen Regelwerk für Schreibregeln und einem Wortschatz, der sowohl Basiswörter als auch Fachwörter beinhalten kann.

An dieser Stelle ist eine Abgrenzung zwischen einer Kontrollierten Sprache, einer Subsprache (Sublanguage) und einer Plansprache hilfreich, um einer Verwechslung vorzubeugen. Wie oben beschrieben ist eine *Kontrollierte Sprache* eine Untermenge ihrer natürlichen Quellsprache, die nach festgelegten Regeln gesteuert bzw. standardisiert wird. Eine *Subsprache* ist zwar eine (gruppen- oder bereichsspezifische) Untergruppe ihrer natürlichen Sprache (Lehrndorfer 1996b: 33), allerdings wird sie *nicht* künstlich gesteuert oder reguliert, sondern „it just happens to have a limited number of linguistic features“ (Roturier 2006: 47). Ein bekanntes Beispiel für eine Subsprache ist die Sprache der Wettervorhersage. Bei

einer *Plansprache* schließlich handelt es sich um eine vollständig künstliche Sprache, wie die bekannte Sprache „Esperanto“ von Zamenhof. Der Wortschatz von Esperanto ist bewusst auf einer Kombination von germanischen, indogermanischen, romanischen und slawischen Wortwurzeln basiert und verfügt über eine einfache phonologische, morphologische und syntaktische Struktur (Lehrndorfer & Reuther 2008: 100). Die Plansprachen zielten auf die internationale Verständigung im Allgemeinen – nicht im technischen Bereich – und die Solidarität der Völker ab, allerdings konnten sie sich aus unterschiedlichen Gründen nicht durchsetzen (Lehrndorfer 1996b: 30). Somit unterscheidet sich eine Plansprache von einer KS nicht nur hinsichtlich des Aufbaus und Umfangs, sondern auch bezüglich des Fokus und der Ziele, denn der technische Bereich stellt herkömmlich den Haupteinsatzbereich der KS dar (vgl. Drewer & Ziegler 2014: 197).

An dieser Stelle darf gleichzeitig nicht unerwähnt bleiben, dass weitere kontrollierte Sprachvarietäten, wie die Einfache bzw. Leichte Sprache, existieren, die in anderen Bereichen wie z. B. bei juristischen, politischen und administrativen Texten eingesetzt werden (Hansen-Schirra & Gutermuth 2018). Die Einfache bzw. Leichte Sprache ist im Kern KS, die anhand von Regeln fachsprachliche Inhalte mit dem Ziel steuern, ihre Komplexität für Menschen mit geringer Sprachkompetenz oder kognitiven Einschränkungen zu reduzieren und somit ihre Verständlichkeit zu erhöhen.[1] Diese Varietäten der KS haben im Zuge der geänderten Rechtslage zur Barrierefreiheit und Inklusion in den letzten Jahren vermehrt an Aufmerksamkeit gewonnen. (ebd.) Die vorliegende Arbeit beschäftigt sich jedoch mit dem herkömmlichen Einsatzbereich der KS, der technischen Dokumentation, daher werden die genauen Ziele der KS im Hinblick auf die technische Kommunikation im folgenden Abschnitt detaillierter beleuchtet.

2.2.2 Ziele der Kontrollierten Sprache

Die Komplexität der natürlichen Sprachen war das Motiv hinter der Entstehung der Kontrollierten Sprachen. Im Bereich der technischen Kommunikation kommt hinzu, dass die technische Dokumentation ein hohes Maß an Verständlichkeit fordert, das sich durch eine natürliche Sprache schwer realisieren lässt. Mithilfe der Kontrollierten Sprachen will man der *Mehrdeutigkeit bzw. Ambiguität* sowohl auf lexikalischer als auch auf syntaktischer Ebene *entgegenwirken*. Konkret

[1] Eine ausführliche Darstellung der kontrollierten Sprachvarietäten, die Einfache und Leichte Sprache, ist u. a. im Duden der Leichten Sprache von Bredel & Maaß (2016) zu finden. Ihre Regeln, die vom „Netzwerk Leichte Sprache“ entworfen wurden, sind unter http://www.bmas.de/SharedDocs/Downloads/DE/PDF-Publikationen/a752-ratgeber-leichte-sprache.pdf?blob=publicationFile&v=2verfügbar.

werden die Kontrollierten Sprachen eingesetzt, um die herkömmlichen Verständlichkeitshindernisse zu bewältigen, die Glasenapp schon 1972 im Zusammenhang mit „Caterpillar Fundamental English“ nannte (Lehrndorfer 1996b: 40f.):

- Lange bzw. verschachtelte Sätze
- Häufung von Nominalphrasen bzw. Aneinanderreihung von Adjektiven in einem Satz
- Unterschiedliche bzw. inkonsistente Satzstrukturen
- Komplexe und zusammengesetzte Zeitformen sowie unregelmäßige Verben
- Inkonsequente Nomenklatur
- Verkürzungen, Ellipsen und Kontraktionen
- Inkonsequente oder falsche Zeichensetzung

Neben der *Verständlichkeit* ist der Gebrauch der KS eine Möglichkeit, „gute *Übersetzbarkeit* von technischer Dokumentation sicherzustellen“ (Lehrndorfer 1996a: 339). Die Verständlichkeit und Übersetzbarkeit waren bereits die angestrebten Ziele von Caterpillar, die ersten Kontrollierten Sprachen (CFE bzw. CTE) zu entwickeln (Detail unter §2.3). Auch wenn die Anzahl der Regeln, die in einer KS gezielt für die Übersetzbarkeit bestimmt sind, limitiert sein kann, dient die Überwindung der obengenannten Verständlichkeitshindernisse und die Reduzierung der Ambiguität im Allgemeinen durch die KS nicht nur der Verständlichkeit, sondern auch indirekt der Übersetzbarkeit (vgl. Fiederer & O’Brien 2009: 53).

Eine *prägnante verständliche Satzstruktur* lässt sich vom Humanübersetzer schneller übersetzen; dementsprechend sinken die Übersetzungskosten (Nyberg u. a. 2003: 248; Göpferich 2007b: 485; Göpferich 2008: 366f.; Lehrndorfer & Reuther 2008: 110f.; Drewer & Ziegler 2014: 208). Ebenfalls zeigen vorherige Studien, dass eine einfache Satzstruktur sich von den MÜ-Systemen besser analysieren bzw. weiterverarbeiten lässt; entsprechend sinkt die Fehlerquote im MÜ-Output und die Notwendigkeit des Post-Editing wird minimiert (Nyberg u. a. 2003: 248; Göpferich 2007b: 481, 485; Göpferich 2008: 366f.; Lehrndorfer & Reuther 2008: 110f.). Diese Studien beziehen sich auf MÜ-Systeme der früheren Ansätze (RBMÜ, SMÜ, PBMÜ und HMÜ). Die vorliegende Studie untersucht, ob und inwiefern ein ähnlicher Effekt verschiedener KS-Regeln nach der Entwicklung des jüngsten MÜ-Ansatzes der NMÜ nachweisbar ist. Darüber hinaus sorgt die Standardisierung für eine hohe *Konsistenz*. Mit dem massiven Wachstum des Volumens

der technischen Dokumentation und dem steigenden Bedarf an Übersetzungen in mehrere Sprachen spielt die Konsistenz eine Schlüsselrolle bei der Textproduktion und Übersetzung. Insbesondere beim Einsatz von Authoring-Memory-Systemen und Translation-Memory-Systemen[2] oder MÜ-Systemen führt die hohe Konsistenz zu einer deutlich effizienteren Textproduktion bzw. Übersetzung (Drewer & Ziegler 2014: 206).

Je nachdem welches Ziel der Einsatz der KS verfolgt, wurde über die letzten 20 Jahre zwischen zwei Typen der Kontrollierten Sprachen unterschieden: den Menschenorientierten und den Maschinenorientierten Kontrollierten Sprachen (Huijsen 1998). Die *Menschenorientierten* Kontrollierten Sprachen (HOCL)[3] zielen auf eine hohe Lesbarkeit und Verständlichkeit durch Regeln wie „Formulieren Sie keine Sätze mit mehr als 20 Wörtern.“, „Formulieren Sie eine Idee pro Satz.“ ab. Auf der anderen Seite versuchen die *Maschinenorientierten* Kontrollierten Sprachen (MOCL),[4] die Verarbeitung der Natürlichen Sprache (Natural Language Processing) in Anwendungen wie der Maschinellen Übersetzung (MÜ) zu unterstützen (vgl. Huijsen 1998; Nyberg u. a. 2003: 247). Lehrndorfer (1996b: 13f.) bezeichnet diesen Typ als den *Maschineneffizienten Ansatz*, der die MÜ oder andere automatische Verarbeitung von technischen Dokumentationen an Effizienz gewinnen lasse. Beispiele für Regeln der MOCL sind (tekom 2013: 137) „Artikel verwenden“ und „Einen Satz nicht durch eine Liste unterbrechen“. Eine gewisse Überschneidung der Regeln beider Typen ist sicherlich vorhanden. Regeln, die die Anzahl der Wörter pro Satz z. B. auf 20 Wörter limitieren oder für eine einfache Struktur sorgen, sind für beide Typen relevant.

2.2.3 Aufbau einer Kontrollierten Sprache

Der Aufbau einer KS erfolgt nach einem der folgenden Ansätze (Reuther 2007; Lehrndorfer & Reuther 2008: 107): dem Präskriptiven Ansatz oder dem Proskriptiven Ansatz. Gleichzeitig findet man in der Praxis eine Mischung aus beiden im Einsatz (Lehrndorfer & Reuther 2008: 108).

Bei dem *Präskriptiven Ansatz* werden Positivregeln mit allen zulässigen grammatischen Strukturen und der zulässigen Lexik vorgegeben (ebd.: 107). Formu-

[2] Während Translation-Memory-Systeme (TMS) für die Erstellung des Zieltexts im Übersetzungsprozess verwendet werden, kommen Authoring-Memory-Systeme (AMS) bei der Erstellung des Ausgangstexts zum Einsatz. Ein AMS speichert die Textsegmente des Ausgangstexts ab und stellt sie zur Wiederverwendung bei der Erstellung neuer Texte zur Verfügung. (Drewer & Schmitz 2017: 196)

[3] Die gängige englische Abkürzung für „Human-oriented Controlled Language“.

[4] Die gängige englische Abkürzung für „Machine-oriented Controlled Language“.

liert der Textproduzent einen Satz, der nicht den vorgegebenen Strukturen entspricht oder ein unzulässiges Wort enthält, wird dieser Satz (meistens von einem CL-Checker, siehe §2.6) markiert. Für den Textproduzenten bedeutet dies, dass jede nicht vorgegebene Struktur bzw. jedes nicht vorgegebene Wort nicht verwendet werden darf. Demzufolge hängt der Einschränkungsgrad beim Schreiben vom Umfang der Positivregeln ab. Im Gegensatz dazu werden bei dem *Proskriptiven Ansatz* Negativregeln mit allen unzulässigen grammatischen Strukturen und Wörtern festgelegt (ebd.). Alle grammatischen Strukturen und Wörter außerhalb dieser Negativregeln sind zulässig. Es besteht somit die Gefahr, dass ein falsches Wort oder eine nicht ideale Struktur ungeprüft bleibt, wenn dies nicht in einer der Negativregeln miteinbezogen wurde. Dementsprechend hängt hierbei der Effizienzgrad vom Umfang der Negativregeln ab.

Nach diesem Vergleich wird deutlicher, dass – obwohl beide Ansätze zu ähnlichen Ergebnissen führen – jeder Ansatz seine Vor- und Nachteile hat (Reuther 2007; Lehrndorfer & Reuther 2008: 107f.): Der Effizienz- und Konsistenzgrad des präskriptiven Ansatzes ist höher, dennoch hat er die Nachteile eines hohen Lern- bzw. Schulungsaufwands und entsprechend niedrigerer Akzeptanz sowie unspezifischer Fehlererläuterung. Auf der anderen Seite verwandelt der proskriptive Ansatz die Nachteile des präskriptiven Ansatzes in Vorteile, indem er den Autoren spezifische Fehlererläuterung sowie größere Formulierungsfreiheit bietet und somit auf höhere Akzeptanz stößt. Zugleich ist die Textvarianz des proskriptiven Ansatzes höher und somit zeigt er sich weniger effizient.

In der vorliegenden Studie werden Regeln aus der Leitlinie „Regelbasiertes Schreiben, Deutsch für die Technische Kommunikation“ des tekom (2013) untersucht. Die tekom-Leitlinie umfasst eine Mischung aus 167 Positiv- und Negativregeln (ebd.), die dem technischen Redakteur jeweils mit einer Entscheidungshilfe zur Verfügung gestellt werden (mehr dazu unter §2.4).

2.3 Entwicklung der Kontrollierten Sprache

In ihrer Forschungsarbeit „The Meaning of Meaning“ ermittelten Ogden & Richards (1923) eine begrenzte Anzahl von Wörtern, die in Wortdefinitionen ständig wiederkehren. Auf Basis dieser Ergebnisse entwickelte Ogden (1935) in den 30er Jahren die erste Kontrollierte Sprache „BASIC English“ (British, American, Scientific, International, Commercial), die aus 850 Wörtern und eingeschränkter Grammatik bestand. Hauptziel damals war, weltweit so vielen Nicht-Englischmuttersprachlern wie möglich zu ermöglichen, Englisch in kürzester Zeit zu lernen und zu benutzen (Schwitter 2007).

Aufbauend auf dem BASIC English wurde ca. 40 Jahre später (1972) im Bereich der technischen Kommunikation die erste firmeninterne KS in Caterpillar Inc. (1974) „Caterpillar Fundamental English" (CFE) entworfen. Die Grundidee bestand darin, dass einerseits stark vereinfachtes Englisch in kurzer Zeit von Englisch-Nicht-Muttersprachlern erlernt und verstanden werden kann und andererseits zusammen mit dem Einsatz vieler Abbildungen Textproduktion und Übersetzungskosten eingespart werden (Drewer & Ziegler 2014: 212). Jedoch stieß das CFE auf Probleme (ebd.), wie die hohe Textkomplexität, die mit einer nur auf 850 Wörter begrenzten KS schwer zu handhaben ist. Dies und die mangelnde Mitarbeiterschulung für die Umsetzung des CFE waren mit einer steigenden Ablehnung der CFE durch die Autoren gekoppelt (ebd.). Demzufolge war die Textqualität so niedrig, dass die Lesbarkeit der Dokumentation darunter litt (Kamprath u. a. 1998). Angesichts dieser Schwierigkeiten in Zusammenhang mit den wachsenden Märkten von Caterpillar war es unmöglich, alle Märkte mit einer vereinfachten Sprache zu bedienen (ebd.). Daraufhin wurde Mitte der 80er Jahre das CFE zu CTE „Caterpillar Technical English" weiterentwickelt (ebd.). Im Fokus des CTE standen eine hohe Textqualität sowie die Erhöhung der Übersetzungsqualität und Reduzierung der Übersetzungskosten (ebd.). Dies sollte damals durch einen höheren Automatisierungsgrad – dank der damaligen schnellen Entwicklung von Soft- und Hardware – erreicht werden (Drewer & Ziegler 2014: 212). Außerdem bestand der Wortschatz des CTE aus ca. 70.000 Wörtern, davon waren mehrere zehntausend Fachtermini (ebd.: 212f.). Zudem war Caterpillar bestrebt den Autoren große Freiheit beim Schreiben zu bieten (ebd.: 213). In der Tat ermöglichte die KS (CTE) Caterpillar Anfang der 90er Jahren sein Textproduktionsvolumen an technischen Dokumentationen erheblich zu steigern und in mehr als 30 Sprachen zu übersetzen (ebd.: 212).

Angesichts dieses Erfolgs begannen in den folgenden Jahren weitere Organisationen firmen- und branchenspezifische KS zu entwickeln, die ihre spezifischen Bedürfnisse berücksichtigen (vgl. Schwanke 1991: 42), wie z. B. das „NCR Fundamental English", das „Plain English Program" (PEP), die „International Language for Servicing and Maintenance" (ILSAM) und das „Perkins Approved Clear English" (Douglas & Hurst 1996: 93), die auf Basis der CFE entwickelt wurden (Schwanke 1991: 42). ILSAM wiederum war die Basis für die Entwicklung des „Simplified Englisch" der AECMA, einer stark kontrollierten Sprache des Englischen, die im Flugzeugwartungsbereich sehr breite Anwendung fand (Göpferich 2008: 370). Darüber hinaus fand die KS Anklang bei multinationalen Konzernen, wie Kodak (Kodak International Service Language), Xerox (Xerox Multilingual Customized English), Sun (Sun Controlled English), Attempto (Attempto Controlled English), Ericsson (Ericsson English), Nortel (Nortel Standard English) und KANT (KANT Controlled English) (Drewer & Ziegler 2014: 211).

Nicht nur im Englischen, auch in anderen Sprachen erschienen in den 90er Jahren Kontrollierte Sprachen, z. B. (ebd.) für Französisch das „Français Rationalisé“, für Spanisch das „Simplified Technical Spanish“ (STS), für Schwedisch das „Scania Swedish“, für Chinesisch das „Controlled Chinese“ und für Japanisch das „Plain Japanese“. Ebenfalls existieren für die deutsche Sprache einige Regulierungsansätze. Da ausgewählte KS-Regeln der deutschen Sprache der Untersuchungsgegenstand dieser Studie sind, wird das Kontrollierte Deutsch im Folgenden näher betrachtet.

2.4 Kontrolliertes Deutsch

Wie die Definition der Kontrollierten Sprache (KS) (unter §2.2.1) darlegt, besteht eine vollständige KS aus einem „Regelwerk mit festen Schreibregeln (v. a. Vorgaben zur Formulierung von Sätzen) sowie vorgegebene[m] Wortschatz aus Basiswortschatz und Fachwortschatz in Form eines Lexikons“ (Drewer & Ziegler 2014: 192). Für die deutsche Sprache existieren einige Ansätze zur Sprachregulierung sowohl im akademischen Bereich als auch in der Industrie. Diese Ansätze basieren jedoch auf Regelwerken ohne vorgegebene Lexika und somit sind sie nach der Standarddefinition der KS nicht als konventionelle vollständige KS zu betrachten. Dies stellt einen wesentlichen Unterschied zwischen englischen KS und dem Kontrollierten Deutsch dar, der auf die folgenden Gründe zurückgeführt wird: Englisch ist die größte Geschäfts- und Forschungssprache. Es ist unbestritten, dass die Anzahl der Englisch-Nicht-Muttersprachler, die leicht erlernbares und verständliches Englisch in ihrem Alltag benötigen, sehr groß ist. Dies ist absolut nicht der Fall bei der deutschen Sprache. Unabhängig von der Anzahl der Muttersprachler zeigt eine Statistik des Instituts SIL International aus dem Jahr 2019, dass Englisch von 753 Mio. Personen, gefolgt vom Chinesischen auf Platz zwei von 199 Mio. und vom Deutschen auf Platz 13 mit nur 38 Mio., als Zweitsprache gesprochen bzw. als erste Fremdsprache gelernt wird.[5] Dementsprechend ist der fehlende Bedarf einer der primären Gründe, warum bisher keine vollständige deutsche KS existiert (ebd.: 216). Der zweite Grund ist linguistischer Natur: Die deutsche Sprache ist im Vergleich zu der englischen Sprache im Hinblick auf den Satzbau sowie die zahlreichen Flexions- und Wortbildungsmorpheme viel komplexer (Lehrndorfer & Reuther 2008: 106). Diese Eigenschaften machen die deutsche Sprache schwer regulierbar. Vor diesem Hintergrund

[5] https://de.wikipedia.org/wiki/Liste_der_meistgesprochenen_Sprachen [abgerufen am 03.08.2019]

haben in der Regel die kontrollierten Varianten der deutschen Sprache die Verbesserung der Verständlichkeit und Konsistenz als Hauptziel (Drewer & Ziegler 2014: 217). Dieses Ziel wird angesichts der Komplexität der Sprache durch grammatische und syntaktische Regeln zusammen mit terminologischen Vorgaben ohne die Festlegung eines kontrollierten Lexikons erreicht (ebd.). Hier wirft sich die Frage auf, ob die Festlegung eines Lexikons nicht erforderlich ist.

Zu den bekanntesten Ansätzen der deutschen KS zählt im akademischen Bereich die Dissertation von Lehrndorfer (1996b) mit dem Titel „Kontrolliertes Deutsch". In ihrer Arbeit geht Lehrndorfer auf die Frage der Notwendigkeit eines Lexikons ein. Sie berücksichtigte die Schwierigkeiten, die mit einer lexikalischen Kontrolle verbunden sind, wie die schwere Lernbarkeit von umfangreichen Lexika, die fehlende Ausdruckmöglichkeit und die inhaltliche Fixierung auf vorgegebene Themenbereiche (ebd.: 139), die wiederum zur geringen Akzeptanz unter den Redakteuren beitragen (ebd.: 137). Demzufolge befürwortet Lehrndorfer, den Fokus bei der Sprachkontrolle eher auf die syntaktische Ebene zu verlagern, wobei das Lexikon durch die kontrollierte Syntax indirekt kontrolliert wird, z. B. durch Regeln wie das Vermeiden von Partizipialkonstruktionen und Funktionsverbgefügen (ebd.: 138).

Mit dem Kontrollierten Deutsch verfolgt Lehrndorfer zwei Ziele (Drewer & Ziegler 2014: 216): (1) die Verbesserung der Lesbarkeit, wobei die Zielgruppe Muttersprachler des Deutschen sind; anders als in der englischen Kontrollierten Sprache, für die Nicht-Muttersprachler die Zielgruppe bilden, und (2) die Verbesserung des Outputs der MÜ. In ihrer Studie wird das Transfer-MÜ-System METAL der Firma Siemens in Kombination mit kontrolliertem Deutsch verwendet. Nach Lehrndorfer (1996b: 155) erfolgt die Sprachkontrolle des Ausgangstexts zunächst durch die Unterscheidung des Redakteurs zwischen drei Aussageabsichten: Handlungsanweisung, Sicherheitshinweis und Aussage zum Produkt (z. B. Produktbeschreibung). Der deutsche Text wird dann auf Konformität mit den KS-Regeln mithilfe der CL-Checker-Funktion[6] des MÜ-Systems METAL geprüft und anschließend mit dem System übersetzt.

In der Industrie ist das „Siemens-Dokumentationsdeutsch" (SDD) das erste firmenspezifische Kontrollierte Deutsch (Göpferich 2008: 375). Das SDD besteht aus Regeln zur Regulierung grammatischer Konstruktionen sowie linguistischer Eigenschaften wie z. B. der syntaktischen Ambiguität (Rascu 2006). Das primäre Ziel der Entwicklung des SDD war nicht die Optimierung der Verständlichkeit durch die Erstellung einfacher Texte – wie der Fall bei der englischen KS (z. B. Caterpillars CFE) –, sondern die Steigerung der maschinellen Übersetzbarkeit

[6]Detail zu CL-Checkern unter §2.6.

mit dem MÜ-System TopTrans von Siemens (Lehrndorfer & Schachtl 1998). Im Übersetzungsprozess prüft TopTrans zunächst, ob die Regeln im Ausgangstext eingehalten wurden (ebd.). Verstöße werden für den Redakteur markiert, damit er sie mit einer interaktiven Unterstützung vom System behebt (Pre-Editing). Nach Abschluss des Pre-Editing wird der Text von TopTrans übersetzt. Auf diese Weise war in der Regel kein Post-Editing mehr erforderlich (ebd.). Dennoch ist das SDD – im Gegensatz z. B. zum Simplified English der AECMA – als eine firmenspezifische Kontrollierte Sprache geblieben und hat sich nicht zu einem Branchenstandard entwickelt (Göpferich 2008: 375). Siemens verfasst weiterhin seine Dokumentationen in Dokumentationsdeutsch, allerdings sind keine aktuellen Veröffentlichungen zu finden, die zeigen würden, wie sich das SDD in den letzten Jahren entwickelt hat. Nicht nur Siemens, sondern auch weitere deutsche Großunternehmen überlassen ihren öffentlichen Auftritt nicht allein dem Geschick der Mitarbeiter (vgl. Baumert & Verhein-Jarren 2012: 153); sie verfassen ihre Dokumentation nach festgelegten Regeln bzw. bestimmten Regelwerken, jedoch erfolgen dazu aus Datenschutz- und Konkurrenzgründen keine Veröffentlichungen.

Ein weiterer praxisbezogener Ansatz zur Steuerung der deutschen Sprache wurde von der tekom Deutschland e. V.[7] in Form einer branchenübergreifenden Leitlinie für die technische Dokumentation entwickelt. Die Regeln, die im Rahmen der vorliegenden Studie analysiert wurden, stammen aus der tekom-Leitlinie,[8] daher wird diese im folgenden Abschnitt genauer betrachtet.

2.4.1 Die tekom-Leitlinie

Als praxisnaher Ansatz entwickelte die Gesellschaft für Technische Kommunikation – tekom (2013) die Leitlinie „Regelbasiertes Schreiben, Deutsch für die Technische Kommunikation“. Anders als eine KS, stellt die Leitlinie einen branchenübergreifenden Standard für die technische Dokumentation mit einer umfangreichen Regelsammlung ohne vorgegebene kontrollierte Lexika dar. Die enthaltenen Regeln kann das Unternehmen u. a. je nach Branche, Zielgruppe und Informationsart an seinen unternehmensspezifischen Bedarf anpassen (Baumert & Verhein-Jarren 2012: 152). Bei der tekom-Leitlinie steht die Steigerung der Qualität und die Reduzierung der Kosten im Dokumentations- und Übersetzungsprozess im Mittelpunkt (Drewer & Ziegler 2014: 217f.). Konkret richtete die tekom-Arbeitsgruppe ihr Augenmerk bei der Erstellung der Leitlinie auf fünf Nutzenzie-

[7] https://www.tekom.de

[8] Die Verwendung der tekom-Leitlinie sowie die Auswahl der analysierten Regeln sind unter §4.5.2.1 begründet.

le: Verständlichkeit, Wiederverwendbarkeit, Konsistenz, übersetzungsgerechtes Schreiben und Überprüfbarkeit (tekom 2013: 9).

Mittlerweile sind die Leitregeln der tekom sowohl in der Forschung als auch in der Industrie weitestgehend etabliert.[9] Dank einer engen Zusammenarbeit zwischen Experten von Hochschulen, der Industrie, Dienstleistungsunternehmen sowie Softwarefirmen bieten die tekom-Regeln ein umfassendes Regelwerk auf sämtlichen Sprach- und Dokumentationsebenen (siehe Tabelle 2.1). Aus diesen Gründen stammen die KS-Regeln, die im Rahmen der vorliegenden Arbeit analysiert wurden, aus der tekom-Leitlinie 2013. Im §4.5.2 werden die in der vorliegenden Studie analysierten Regeln detailliert präsentiert und ihre Auswahl begründet. Im Folgenden werden die Entwicklung sowie der Aufbau der tekom-Leitlinie näher aufgegriffen.

Der ursprüngliche Arbeitskreis für die Erstellung der Leitlinie strebte einen funktionsorientierten Textaufbau, stilistische Vorgaben auf Satzebene sowie terminologische Vorgaben an (Drewer & Ziegler 2014: 218). Die Leitlinie sollte Regeln zur Benennungsbildung, d. h. Festlegung von Benennungen für neue Begriffe (vgl. Drewer & Schmitz 2017: 70ff), sowie ein Lexikon mit branchenübergreifendem Vokabular beinhalten (Drewer & Ziegler 2014: 218). Die Festlegung von Regeln zur Benennungsbildung ist keine Neuheit, hingegen würde die Entwicklung eines Lexikons mit branchenübergreifendem Vokabular zu einer traditionellen KS führen. Nach einem Jahr wurde aus dem Arbeitskreis eine tekom-Arbeitsgruppe gegründet, die die Arbeit an den Regeln fortgesetzt hat. (ebd.)

Im Jahr 2011 wurde die erste Auflage der Leitlinie von der Arbeitsgruppe „Technisches Deutsch“ entworfen (tekom 2013: 13). Die Regeln sollten eine „Leitlinie für professionelles Deutsch in der Technischen Kommunikation“ bilden (Drewer & Ziegler 2014: 218). Inhalt der „Leitlinie Technisches Deutsch“ war ein „Regelsatz, der z. B. Vorschriften zur Verwendung bestimmter Wortformen, zur Wortbildung, zum Satzbau, zur Bildung von Abkürzungen, zur Zeichensetzung, [sic] sowie Regeln zur Festlegung von Benennungen“ umfasste (tekom 2009). Wie der Inhalt zeigt, liegt der Fokus auf dem Satzbau und der Terminologie, nicht auf der Etablierung eines vollständigen Vokabulars. Somit bietet die tekom-Leitlinie eine Sammlung von Schreib- und Stilregeln und keine klassische KS (Drewer & Ziegler 2014: 218).

Im Sommer 2010 führte die Arbeitsgruppe einen Beta-Test durch, um die Anwendbarkeit und den Praxisbezug der Leitlinie vor der Veröffentlichung zu prü-

[9] Die Tekom-Regeln sind der Kernregelsatz in zwei marktführenden CL-Checkern Acrolinx (https://www.acrolinx.de/produktuberblick/) und CLAT (vgl. Geldbach 2009). Mehr zum CLAT und seiner Entwicklung unter §2.6.

fen. 39 Tester (Anfänger, Fortgeschrittene und Experten) haben die Leitlinie bewertet und die Ergebnisse flossen bei der Entwicklung der ersten Auflage ein (Fleury o. D.).[10] Daraufhin erschien die erste Auflage der Leitlinie 2011 und zwei Jahre später wurde die zweite Auflage veröffentlicht (2013). In der Leitlinie werden die Regeln wie folgt dargestellt:

- Regelnummer: besteht aus einem Buchstaben und einer dreistelligen Zahl
- Regelüberschrift
- Regelbeschreibung
- Handlungsanweisungen: zeigen, wie die Regel umgesetzt werden kann zusammen mit Tipps für die Umsetzung
- Negative und positive Beispiele für die Umsetzung der Regel tabellarisch dargestellt
- Entscheidungshilfe: erklärt, wann der Einsatz der Regel empfohlen ist und zeigt Alternative auf
- Maschinelle Prüfbarkeit: gibt an, ob das Einhalten der Regel maschinell (d. h. mithilfe von Controlled-Language-Checkern) geprüft werden kann

Die zweite Auflage ist wie folgt aufgebaut (tekom 2013):

Wie Tabelle 2.1 zeigt, deckt die tekom-Leitlinie mit insgesamt 167 Regeln sämtliche Sprach- und Dokumentebenen in der technischen Dokumentation umfassend ab (tekom 2013). Aus diesen Regeln kann das Unternehmen individuell auswählen und somit seine Unternehmenssprache (Corporate Language) auf eine Weise gestalten, die es ihm ermöglicht, seine Unternehmensidentität (Corporate Identity) auszudrücken. So können Unternehmen derselben Branche Ihre Dokumentation nach unterschiedlichen Regeln erstellen und dabei einen unterschiedlichen Grad an Kundenorientierung (z. B. Wert auf Kundenzufriedenheit) zum Ausdruck bringen.

Einer Ablehnung seitens der technischen Redakteure wird z. B. mittels der Bereitstellung von einer Entscheidungshilfe bei jeder Regel entgegengewirkt. Beispielsweise erklären die Entscheidungshilfen bei den Regeln zur Vermeidung vom Passiv, dass die Verwendung vom Passiv sinnvoll sein kann, „wenn der Handelnde nicht bekannt ist oder bewusst nicht genannt werden soll“ (tekom 2013:

[10] Angaben von Fleury (o. D.) tekom-Vorstandspatin der Arbeitsgruppe.

Tabelle 2.1: Übersicht des Aufbaus der tekom-Regeln. Quelle: tekom 2013

Anzahl	Regelebene / Kategorie Unterebene / Unterkategorie	Beispiel	
40	**1. Regeln auf Textebene**		
29	Dokumentstruktur	T 105	Redundanzen in Überschriften vermeiden
11	Informationsstruktur	S 306	Aufzählungen als Liste darstellen
41	**2. Regeln auf Satzebene**		
7	Vermeidung von mehrdeutigen Konstruktionen	S 102	Eindeutige pronominale Bezüge verwenden
4	Vermeidung von unvollständigen Konstruktionen	S 204	Keine Wortteile weglassen
13	Vermeidung von komplexen Konstruktionen	S 303	Partizipialkonstruktionen vermeiden
2	Wortstellung und Abfolge von Satzelementen	S 402	Eingeschobene Nebensätze vermeiden
15	Stilistische Regeln	S 501	Vorgangspassiv vermeiden
50	**3. Regeln auf Wortebene**		
28	Wortbildung	B 108	Komposita mit Ziffern immer mit Bindestrich
3	Abkürzung	B 203	Abkürzungsschreibweisen festlegen
4	Verwendung von Benennungen und Zahlen	B 302a	Zahlen von 1 bis 12 in Ziffern schreiben
14	Lexikalische Vorgaben	L 103	Funktionsverbgefüge vermeiden
1	**4. Rechtschreibung**	R 101	Einheitlichen Rechtsschreibstil verwenden
21	**5. Zeichensetzung**	Z 103b	Für zitierte Oberflächentexte gerade Anführungszeichen "..." verwenden
6	**6. Platzsparendes Schreiben**	P 104	Wörter konsistent und nachvollziehbar kürzen
8	**7. Übersetzungsgerechtes Schreiben**	Ü 103	Sinneinheiten mit einem Punkt oder Absatzwechsel abschließen

80). Auf der anderen Seite empfiehlt die Entscheidungshilfe die Verwendung des Aktivs z. B. bei der Formulierung von handlungsorientierten Informationseinheiten, da das Aktiv verdeutlicht, „wer eine Handlung ausführt oder ausführen soll […] und motiviert den Leser die Anleitung zu befolgen“ (tekom 2013: 81).

Die Parallelen zwischen dem Kontrollierten Deutsch von Lehrndorfer (1996b) und der tekom-Leitlinie 2013 lassen sich deutlich erkennen: Beide Werke bestehen aus Regelsätzen und umfassen kein Lexikon. Das bietet insbesondere für eine Sprache mit reicher Morphologie wie Deutsch einen guten Ausweg aus der Problematik der geringen Akzeptanz bzw. der psychischen Abneigung der KS durch die Textproduzenten (vgl. Nyberg u. a. 2003: 249; Drewer & Ziegler 2014: 209) an. Diese Problematik wird bekanntlich oft in Zusammenhang mit der englischen KS genannt, da das Beachten eines Lexikons einschränkend ist, viel Vorsicht beim Schreiben erfordert und das Schreiben (zumindest in der Anfangsphase der Arbeit mit der KS) verlangsamen kann.

Hinsichtlich der (maschinellen) Übersetzbarkeit ist anschließend Folgendes anzumerken: Obwohl auf den ersten Blick in Tabelle 2.1 nur Kategorie 7 „Übersetzungsgerechtes Schreiben“ mit acht Regeln gesondert für die Übersetzbarkeit zu sehen ist, zeigen weitere Regeln eine Wirkung auf die (maschinelle) Übersetzbarkeit. Diese Wirkung wurde von weiteren zitierten Studien (vgl. Bernth & Gdaniec 2001; Reuther 2003; Siegel 2011; Congree 2018) in Zusammenhang mit den früheren MÜ-Ansätzen untersucht (siehe Überblick in §4.5.2.2). Die Grundidee bei diesen Studien ist, dass die KS im Allgemeinen die Ambiguität und die Satzkomplexität reduziert sowie die Satzstruktur vereinfacht; und auf diesem Wege indirekt zur Verbesserung der (maschinellen) Übersetzbarkeit beiträgt.

Nach dieser Darstellung des Kontrollierten Deutsch und seiner Besonderheiten sollen im Folgenden die Stärken und Schwächen der KS aus unterschiedlichen Perspektiven näher beleuchtet werden.[11]

2.5 Stärken und Schwächen der Kontrollierten Sprache

Der Einsatz der KS ist mit vielen Stärken verbunden, gleichzeitig dürfen ihre Schwächen nicht außer Acht gelassen werden. Je nachdem in welchem Umfang die KS eingesetzt wird, angefangen beim Einsatz einer vollständigen KS bis zum

[11] Das Institut für technische Literatur (itl AG) bietet in einem aktuellen Leitfaden (Stand: November 2019) eine Übersicht der 14 wichtigsten Normen und Richtlinien für die technische Dokumentation, darunter die hier präsentierte Leitlinie der tekom. In dem Leitfaden werden die Normen zusammengefasst und kommentiert. Außerdem wird erläutert, warum sie wichtig sind. Näheres dazu unter: https://www.itl.eu/de/nachrichten/details/der-itl-normenguide.html.

Einsatz einer Redaktionsleitlinie, können einige Stärken bzw. Schwächen entfallen bzw. in ihrer Gewichtung variieren. Im Folgenden werden die Stärken und Schwächen nach den vier Stakeholdern, also dem Unternehmen, dem Redakteur, dem Übersetzer und dem Rezipienten gegenübergestellt und erörtert:

2.5.1 Für die Unternehmen

Erstens bildet das Unternehmen durch das Einhalten von vordefinierten lexikalischen, syntaktischen und stilistischen Regeln eine Unternehmenssprache (Corporate Language), die einen Bestandteil seiner Unternehmensidentität (Corporate Identity) darstellt (vgl. Lehrndorfer & Reuther 2008: 111; Drewer & Ziegler 2014: 206). Zweitens erhöhen die Lesbarkeit und Verständlichkeit – als Teil der Hauptziele der KS – (vgl. Nyberg u. a. 2003: 248; Lehrndorfer & Reuther 2008: 110) die Kundenzufriedenheit und verringern gleichzeitig die Haftungsansprüche (Drewer & Ziegler 2014: 206). Drittens verbessert der Einsatz einer KS die Qualität der Dokumentation und beschleunigt ihre Übersetzung (Nyberg u. a. 2003: 255). Dies reduziert die Übersetzungskosten, ermöglicht den Unternehmen fremdsprachige Handbücher schneller zu erstellen und verkürzt somit die Time-to-Market der zugehörigen Produkte. Eine kürzere Time-to-Market fördert wiederum die Konkurrenzfähigkeit des Unternehmens. (ebd.)

Gleichzeitig ist die Entwicklung und Implementierung einer KS mit Zeit- und Kostenaufwand verbunden (vgl. Nyberg u. a. 2003: 249): Die KS wird entweder vom Unternehmen entwickelt, gestaltet und verwaltet oder eine vorhandene KS wird lizenziert und nach Unternehmensbedürfnissen angepasst. Die Einführung einer KS umfasst mehrere Phasen der linguistischen Analyse, den Terminologieaufbau bzw. die Entwicklung oder den Kauf eines CL-Checkers. Ferner muss die KS nach der Einführung gepflegt und kontinuierlich zusammen mit der Terminologie an die aktuellen Standards und Bedürfnisse des Unternehmens angepasst werden. (ebd.) Zudem muss die Lern- und Umgewöhnungsphase der Redakteure und die während dieser Phase ggf. noch nicht hohe Qualität in Kauf genommen werden (Drewer & Ziegler 2014: 209).

2.5.2 Für die Redakteure

In der technischen Kommunikation verfassen die Redakteure viele Dokumentationen, in denen zahlreiche Sätze sich vollständig oder teilweise wiederholen; Beispiele hierfür sind Installationsanweisungen bei Software-Updates, Bedienungsanleitungen von Geräten, Montageanweisungen von Maschinen sowie Allgemeine Geschäftsbedienungen. Durch die Einhaltung von Regeln der KS – insbesondere bei dieser Art von Dokumentationen – wird sichergestellt, dass sowohl die

verwendeten Terminologien als auch Formulierungen konsistent sind. Das wird in der Regel durch den Einsatz der KS zusammen mit der Verwendung eines Authoring-Memory-Systems[12] oder CL-Checkers[13] realisiert. Verfasst der Redakteur mithilfe eines Authoring-Memory-Systems einen Satz, der in dem aktuellen oder einem anderen Dokument vollständig (100%-Match) oder zum Teil (Fuzzy-Match) vorkam, zeigt das System diesen Satz an (Drewer & Schmitz 2017: 196). Der Redakteur hat dann die Möglichkeit, den angezeigten Satz zu übernehmen und ggf. anzupassen (Drewer & Schmitz 2017: 196). Diese Möglichkeit erleichtert und beschleunigt die Arbeit des Redakteurs und steigert die Qualität durch die hohe terminologische und stilistische Konsistenz (vgl. Göpferich 2008: 367; Drewer & Schmitz 2017: 197). Diese Konsistenz ist multiplizierbar, denn sollten mehrere Redakteure an einem Dokument gleichzeitig oder an mehreren Dokumenten über einen längeren Zeitraum arbeiten, bietet die KS eine enorme Unterstützung für die sprachliche und stilistische Vereinheitlichung, die durch den Einsatz von Software sichergestellt wird (vgl. Göpferich 2008: 367; Drewer & Ziegler 2014: 235). Ferner sehen einige Autoren in der KS einen Orientierungsrahmen, der sie bei einer schnellen Entscheidung über Formulierungen unterstützt (Drewer & Ziegler 2014: 207).

Auf der anderen Seite können die Textproduzenten in manchen Fällen Fachspezialisten sein, die über ein begrenztes linguistisches Wissen verfügen, z. B. in Kleinunternehmen. Diese Gruppe von Textproduzenten kann Schwierigkeiten haben, linguistische Regeln wie[14] „Funktionsverbgefüge vermeiden“ oder „Komposita aus vier und mehr Basismorphemen immer mit Bindestrich schreiben“ umzusetzen. Eine weitere mögliche Problematik des Einsatzes von KS besteht darin (Nyberg u. a. 2003: 248), dass das Schreiben durch die Einhaltung der KS-Regeln zeitaufwendiger ausfällt, manche Sätze müssen vollständig umformuliert werden. Die Verlangsamung des Schreibprozesses kann zu einer Abneigung gegen die Verwendung der KS führen. Manche Redakteure entwickeln eine psychische Abneigung gegen die Einhaltung der KS-Regeln. Sie finden, dass die KS ihre Kreativität und Motivation hemme, da sie beim Schreiben eingeschränkt seien (vgl. Nyberg u. a. 2003: 249; Lehrndorfer & Reuther 2008: 112; Drewer & Ziegler 2014:

[12] Während Translation-Memory-Systeme (TMS) für die Erstellung des Zieltexts im Übersetzungsprozess verwendet werden, kommen Authoring-Memory-Systeme (AMS) bei der Erstellung des Ausgangstexts zum Einsatz. Ein AMS speichert die Textsegmente des Ausgangstexts ab und stellt sie zur Wiederverwendung bei der Erstellung neuer Texte zur Verfügung. (Drewer & Schmitz 2017: 196)

[13] Ein CL-Checker (Controlled-Language-Checker) ist eine spezielle Software zur Prüfung, ob die KS-Regeln eingehalten werden. Da in der vorliegenden Studie einen CL-Checker verwendet wurde, wird die Funktionsweise dieser Software detailliert unter §2.6 erläutert.

[14] Beispiele aus der tekom 2013.

209). Zudem müssen die Redakteure den Lernaufwand der KS-Regeln und des dafür verwendeten Tools auf sich nehmen (Drewer & Ziegler 2014: 209).

2.5.3 Für die Übersetzer

Erhält der Übersetzer vom technischen Redakteur einen terminologisch und stilistisch konsistenten und eindeutig formulierten Ausgangstext, erleichtert und beschleunigt dies seine Übersetzungsaufgabe, denn die lexikalische und syntaktische Klarheit sowie die Konsistenz erhöhen die Lesbarkeit und die Verständlichkeit für den Übersetzer als Textrezipient (vgl. Nyberg u. a. 2003: 248; Göpferich 2008: 366f.; Drewer & Ziegler 2014: 208). Dies wiederum ermöglicht ihm, die Übersetzung in kürzerer Zeit anzufertigen und spiegelt sich in niedrigeren Übersetzungskosten wider. Ferner werden die Vorteile der KS bei der Übersetzung maximiert, wenn das Unternehmen ein Translation-Memory-System im Einsatz hat. In einem Translation-Memory-System werden vorherige Übersetzungen gespeichert. Erkennt das System einen Ausgangssatz, der vorher im selben oder in einem anderen Dokument vollständig (100%-Match) oder zum Teil (Fuzzy-Match) übersetzt wurde, zeigt es diesen Satz zusammen mit seiner Übersetzung an (vgl. Drewer & Ziegler 2014: 208; Drewer & Schmitz 2017: 213f.). Da das Einhalten von den KS-Regeln die Konsistenz im Ausgangstext erhöht, steigen im Translation-Memory-System die Matchquoten, d. h. die Wiederverwendbarkeit von vorherigen Übersetzungen (Nyberg u. a. 2003: 248; Drewer & Ziegler 2014: 206). Das wiederum resultiert in einer effizienten Übersetzung (Lehrndorfer & Reuther 2008: 111; Drewer & Ziegler 2014: 206).

Im Bereich der maschinellen Übersetzung kamen vorherige Studien zu dem Ergebnis, dass ein Pre-Editing mithilfe der KS-Regeln einen positiven Einfluss auf den MÜ-Output hat, unter anderem im Sinne eines geringen Post-Editing-Aufwands und niedriger Post-Editing-Zeit (vgl. Göpferich 2007b: 481; Göpferich 2008: 367; Lehrndorfer & Reuther 2008: 110). Die Hauptidee dieser Studien lautet wie folgt: Achtet der Autor bei der Textproduktion auf Eindeutigkeit und vermeidet komplexe Satzstrukturen, können viele Fehlerquellen für die Übersetzung eliminiert werden (vgl. Göpferich 2007b: 485). Das System produziert folglich einen besseren Output (Nyberg u. a. 2003: 248), insbesondere wenn die Terminologie zuvor in die Systemdatenbank eingepflegt wurde (Göpferich 2008: 366). Diese Studien wurden für die frühen MÜ-Ansätze (RBMÜ, SMÜ, PBMÜ und HMÜ) durchgeführt. Ob und inwiefern der Einsatz der verschiedenen KS-Regeln nach der Entwicklung des jüngsten MÜ-Ansatzes der NMÜ den MÜ-Output verbessert, ist eine zentrale Fragestellung der vorliegenden Studie.

Diese Vorteile können in manchen Fällen mit Nachteilen für die Übersetzer gekoppelt sein. Die aufgrund der höheren Konsistenz gestiegenen Matchquoten können dazu führen, dass der Umfang der Arbeit für die Übersetzer stark abnimmt. Manche Unternehmen setzen die (englische) KS ein, um die Humanübersetzung von – meistens internen – Dokumenten zu vermeiden (Drewer & Ziegler 2014: 208). Vergleichbar mit den Nachteilen bei den Redakteuren kann die Einhaltung von KS-Regeln auch bei den Übersetzern zeitaufwendiger und komplexer als das Übersetzen ohne den Einsatz von KS-Regeln ausfallen (Nyberg u. a. 2003: 249). Ebenfalls nehmen einige Übersetzer wahr, dass ihre Schreibfähigkeiten durch den Einsatz der KS eingeschränkt sind (ebd.).

2.5.4 Für die Rezipienten

Auf Basis der oben diskutierten Wirkungen der KS lässt sich zusammenfassend schlussfolgern, dass der Rezipient – insbesondere Nicht-Muttersprachler – ebenfalls von dem Einsatz der KS durch die erhöhte Eindeutigkeit, die verbesserte Lesbarkeit und die optimierte Verständlichkeit profitiert (Göpferich 2008: 366; Drewer & Ziegler 2014: 208).

Auf der anderen Seite kann der Stil der KS von manchen Rezipienten als eintönig und unästhetisch empfunden werden, da sie gewöhnlich einen abwechslungsreichen Stil erwarten und diesen als lebendig empfinden (Lehrndorfer & Reuther 2008: 112f.).

Angesichts der obengenannten Stärken und gleichzeitig des Aufwands und der Kosten des Einsatzes einer KS beschäftigten sich mehrere Studien (vgl. Nyberg u. a. 2003: 248; Göpferich 2008: 369; Lehrndorfer & Reuther 2008: 108) mit der Frage, wann es empfohlen ist, eine KS zu implementieren. Die Antwort lässt sich wie folgt zusammenfassen:

- bei einem wachsenden Dokumentationsumfang, insbesondere wenn an der Dokumentation mehrere Personen oder Stellen arbeiten,
- bei einer häufigen Kombinierung von mehreren Dokumenten oder bei einer häufigen Änderung des Inhalts,
- für technische Inhalte insbesondere bei einer steigenden Komplexität des Inhalts bzw. einem zunehmenden Dokumentationsverwaltungsaufwand,
- wenn große Textvolumen in verschiedene Sprachen übersetzt werden,
- wenn die Kosten der Dokumentation bzw. der Übersetzung reduziert werden sollen.

2.5.5 Diskussion der Stärken und Schwächen der KS

Im Folgenden werden die obengenannten Stärken und Schwächen der KS (kritisch) reflektierend diskutiert:

Das Gegenargument, dass der Einsatz von KS mit hohen Kosten verbunden wäre, ist heutzutage nicht nachvollziehbar, denn jedes Unternehmen, das einen professionellen Dokumentations- und Übersetzungsprozess anstrebt, besitzt und pflegt bereits eine Terminologiedatenbank und wendet eine Form der Sprachkontrolle an, sei es in Form von Styleguide oder Redaktionsleitfaden (vgl. Göpferich 2007a). In seinem Aufsatz „Implementing a Controlled Language is now cheaper and easier than ever" nennt Mügge (2013) Beispiele[15] für kommerzielle Regelwerke und leistungsstarke Controlled-Language-Checker, die kostenlos zur Verfügung stehen und somit die Implementierung von KS auch für Kleinunternehmen attraktiv gestalten. Außerdem lässt sich die Kostenersparnis, die das Unternehmen durch den Einsatz von KS realisiert, anhand eines einfachen Rechenbeispiels verdeutlichen: Die technische Dokumentation beinhaltet zahlreiche Standardsätze, die sich in den Dokumentationen wiederholen. Wenn 30 Standardsätze in 5 Dokumenten vorkommen, wobei sie leicht an den Kontext angepasst werden sollen, hätten die Redakteure ohne Einsatz von KS-Regeln 150 Sätze (30 Sätze * 5 Dokumente), die neu formuliert werden müssen. Im Falle des Einsatzes von KS-Regeln müssten die 30 Sätze in den 5 Dokumenten nur an den Kontext angepasst werden. Dies bietet nicht nur eine große Zeitersparnis, sondern auch einen konsistenten Text. Sollten die 5 Dokumente in 3 Fremdsprachen übersetzt werden, steigt die Ersparnis deutlich (450 Sätze ohne KS (30 Sätze * 5 Dokumente * 3 Zielsprachen) im Vergleich zu 90 Sätzen (30 Sätze * 3 Zielsprachen) mit leichten Anpassungen bei dem Einsatz von KS-Regeln). Zudem spiegelt sich die Konsistenz in den Zielsprachen wider.

Bezüglich der Problematik der psychischen Abneigung: Es ist nachvollziehbar, dass jeder neue Prozess im Unternehmen von seiner Zielgruppe erlernt werden muss und dass die Lernphase einen gewissen Aufwand mit sich bringt. Nachdem jedoch ein Lerneffekt verzeichnet wird, geben die Redakteure in manchen Unternehmen an, dass sie eine „Sensibilisierung für sprachliche Sachverhalte und eine effizientere Texterstellung und -bearbeitung" entwickelt haben (Lehrndorfer & Reuther 2008: 111). In der Einführungsphase von KS wird empfohlen, die Redakteure und Übersetzer bei der KS-Definition und dem KS-Einsatz einzubeziehen,

[15] Der „ASD-STE100-Regelsatz" der AeroSpace- und Defense Industries Association of Europe (ASD) sowie der Open-Source-Checker „STE Term Checker", die für ASD-STE100 zur Grammatik, Stil- und Vokabelprüfung optimiert wurde, stehen kostenlos zur Verfügung.

um einer potentiellen Abneigung vorzubeugen (vgl. Nyberg u. a. 2003: 249). Allerdings ist dieser Ansatz in mittelständischen und Großunternehmen durch die hohe Anzahl der Redakteure und Übersetzer nur bedingt realisierbar. Außerdem benötigt jedes Unternehmen in fortgeschrittenen Phasen des KS-Einsatzes sowie mit der Rotation von neuen Redakteuren und Übersetzern eine solide Basis, um eine erreichte KS-Akzeptanz beizubehalten bzw. einer möglichen Abneigung entgegenzuwirken. Da die häufig genannten Schwierigkeiten des KS-Einsatzes auf den erhöhten Zeitaufwand, die erhöhte Schreibkomplexität sowie eingeschränkte Kreativität zurückgeführt werden, stehen den Unternehmen mehrere Wege zur Verfügung, um diese Schwierigkeiten zu bewältigen:

Erstens ist die Verwendung von einem Controlled-Language-Checker (CLC),[16] Authoring-Memory-System und Translation-Memory-System[17] erforderlich, um die Umsetzung der KS-Regeln zu erleichtern und entsprechend den damit verbundenen Zeitaufwand zu minimieren. In der Praxis muss nicht selten ein großes Volumen an technischen Dokumentationen unter Zeitdruck verfasst werden. Mithilfe eines Redaktionstools wird die Arbeit des Redakteurs strukturierter und vereinfacht, denn die Redaktionstools zentralisieren die relevanten Terminologiedaten und Redaktionsregeln vor dem Redakteur auf dem Bildschirm.

Zweitens stellt der Einsatz von einer Mischung aus präskriptiven und proskriptiven Regeln eine Lösung dar, um die Dokumente abhängig davon, wie kritisch sie sind (z. B. Sicherheitsanweisung im Vergleich zu einem internen Mitteilungsdokument) – mithilfe des CLC – mit bestimmten KS-Regeln der beiden Arten zu verknüpfen. Auf diesem Weg kann den Redakteuren Flexibilität und Freiraum für Kreativität durch die proskriptiven Regeln verschafft werden, solange die Natur des Dokuments dies zulässt.

Der dritte Weg ist die Schulung von neuen Mitarbeitern. Schulungen sind zwar mit Kosten verbunden, sie tragen aber im Endeffekt zur Dokumentationsqualität bei, wodurch wiederum finanzielle Nutzen unter anderem durch eine höhere Kundenzufriedenheit, verringerte Haftungsansprüche sowie verbesserte Konkurrenzfähigkeit realisiert werden können.

Bezüglich der Arbeit der Übersetzer kann bei einer reinen Übersetzungsaufgabe der Arbeitsaufwand zwar abnehmen bzw. beim MÜ-Einsatz die Arbeit auf

[16] Ein CLC ist ein Programm, das den technischen Redakteur unterstützt, die Regeln der KS (korrekt) einzusetzen (mehr dazu unter §2.6).

[17] Während Translation-Memory-Systeme (TMS) für die Erstellung des Zieltexts im Übersetzungsprozess verwendet werden, kommen Authoring-Memory-Systeme (AMS) bei der Erstellung des Ausgangstexts zum Einsatz. Ein AMS speichert die Textsegmente des Ausgangstexts ab und stellt sie zur Wiederverwendung bei der Erstellung neuer Texte zur Verfügung. (Drewer & Schmitz 2017: 196)

Post-Editing eingeschränkt werden, dennoch bleibt die Rolle der Übersetzer unentbehrlich und ihre Bedeutsamkeit maximiert sich zusammen mit den mit dem Gegenstand verbundenen Risiken, dessen Dokumentation übersetzt werden soll (z. B. Risiken der Produkthaftung oder der Produktsicherheit). Denn je höher diese Risiken sind, desto umfangreicher ist der Risikobewältigungsaufwand, der letztendlich von den Übersetzern geleistet wird (Canfora & Ottmann 2015).

Abschließend sollte auf ein geläufiges Argument eingegangen werden, dass die Leser eine nach KS-Regeln verfasste Dokumentation als eintönig bzw. nicht abwechslungsreich empfinden. Ein vergleichbarer Effekt der verminderten Akzeptabilität wurde von Hansen-Schirra & Maaß (2020) bei der Leichten Sprache, die als eine Varietät der KS gilt, festgestellt. Gezielt für die technische Dokumentation behandelte Püschel (1996: 335f.) bereits 1996 in seinem Beitrag „Sprachstil – ein Thema für Technische Redakteure?“ diese Frage und empfahl „den Text so abwechslungsreich wie möglich zu machen“, denn „auch ein Stilbruch kann die Aufmerksamkeit wecken“. In der technischen Dokumentation muss der Einfluss der KS auf den Stil je nach Textsorte bzw. Dokumentart differenziert betrachtet werden. Handelt es sich um Dokumentationen, die z. B. Verfahren oder Instruktionen vermitteln, wäre es unangebracht, die KS in dieser Hinsicht zu kritisieren, denn vor allem durch die konsistente Wortwahl und Satzstruktur sowie den einheitlichen Stil in solchen technischen Dokumentationen wird die Lesbarkeit gesteigert, wodurch der Leser Zeit spart und eine bessere Orientierung erhält (vgl. Farkas 1985). Geht es hingegen um Dokumentationen, die zwar technische Angaben, allerdings z. B. für Zwecke des Marketings, vermitteln, kann ein eintöniger Stil fraglich sein.

2.6 Controlled-Language-Checker (CL-Checker)

2.6.1 CL-Checker – Überblick

Die IT-Unterstützung beim Dokumentationsprozess ist heutzutage unerlässlich. Aufgrund der Komplexität des Dokumentationsprozesses und -umfangs ist es insbesondere in großen Unternehmen kaum vorstellbar, auf Systeme der Terminologieverwaltung, Redaktion, Sprachkontrolle bzw. -überprüfung, Übersetzung und des Content-Managements zu verzichten. Im Redaktionsprozess werden in der Regel Controlled-Language-Checker, Authoring-Memory-Systeme oder Content-Management-Systeme verwendet. Ein Content-Management-System (CMS) ist ein umfangreiches Softwaresystem zur Unterstützung des Content Managements. Es besteht i. d. R. aus drei Anwendungsmodulen: einem Redaktionssystem zur

Bearbeitung und Verwaltung von Inhalten, einem Content Repository zur Speicherung der Inhalte und einem Publishing System zur Ausgabe der Inhalte.[18] Ein Authoring-Memory-System (AMS) ist ein Softwaresystem, das die Textsegmente des Ausgangstexts abspeichert und sie zur Wiederverwendung bei der Erstellung neuer Texte zur Verfügung stellt (Drewer & Schmitz 2017: 196). In diesem Abschnitt liegt der Fokus auf der für die Studie relevanten Software, dem Controlled-Language-Checker (CL-Checker).

Ein Controlled-Language-Checker (CLC) ist ein Programm, dessen Ziel darin besteht, die technischen Redakteure zu unterstützen, die Regeln der KS (korrekt) einzusetzen. Es handelt sich hierbei meistens[19] um eine frei konfigurierbare Software, mit der Redaktionsleitlinien oder andere Regelbestände unternehmensspezifisch abgebildet und maschinell geprüft werden können (Drewer & Ziegler 2014: 227). Die Software wird in der Regel mit Standardstil- und Grammatikregeln sowie Regeln zur Rechtschreibung und Zeichensetzung vom Hersteller geliefert (Geldbach 2009). Die eingesetzten Regeln und ihren Umfang kann jedes Unternehmen frei bestimmen. Die Terminologie ist unternehmensspezifisch und wird daher vom erwerbenden Unternehmen individuell eingepflegt. Die hinterlegten Regeln können je nach Zielgruppe, Textsorte und anderen definierbaren Merkmalen angepasst werden (Drewer & Ziegler 2014: 234). Dementsprechend spielen die CLCs eine bedeutende Rolle dabei, den Schwächen der KS entgegenzuwirken, indem ihre Konfigurierbarkeit dem Unternehmen und den Redakteuren mehr Flexibilität und Individualität bietet (vgl. Rösener 2010).

Die Konzeption des Controlled-Language-Checkers hängt von den zwei Ansätzen der KS (siehe §2.2.3) ab (Drewer & Ziegler 2014: 228ff.): Beim *Präskriptiven Ansatz* müssen alle zulässigen Vokabulare sowie syntaktischen Strukturen in dem CLC hinterlegt werden. Sollte der Autor ein Wort oder eine Struktur angeben, das/die nicht im CLC vorliegt, wird dies markiert, damit der Autor es/sie ersetzt. Man kann bei diesem Ansatz keine automatische Korrektur erwarten, da es nicht selten zu fehlerhaften Fehlermeldungen (Noise) kommt. Ob und wie eine Korrektur erfolgt, bleibt die Entscheidung des Autors. Daher ist dieser Ansatz sehr umfangreich und erfordert einen hohen Lern- bzw. Schulungsaufwand. Beim *Proskriptiven Ansatz* hingegen müssen alle unzulässigen Vokabulare sowie syntaktischen Strukturen in dem CLC hinterlegt werden. Der Autor hat entsprechend einen größeren Freiraum im Vergleich zum präskriptiven Ansatz. Die Fehlermeldungen sind spezifisch und von daher gut umsetzbar. All dies führt

[18] https://wirtschaftslexikon.gabler.de/definition/content-management-system-cms-31303

[19] Es gibt auch nicht frei konfigurierbare CLCs, die mit bestimmten Regeln geliefert werden. Die gelieferten Regeln können zwar eingeschränkt, aber nicht um kundenspezifische Regeln erweitert werden (Drewer & Ziegler 2014: 232).

zu einer höheren Akzeptanz bei den Autoren, bietet jedoch aufgrund des großen Freiraums weniger Konsistenz im Vergleich zu dem präskriptiven Ansatz. Aufgrund der Stärken und Schwächen der beiden Ansätze werden sie auch kombiniert eingesetzt (ebd.: 230).

2.6.2 Die Software und ihre Funktionsweise

CLCs haben eine Client-Server-Architektur,[20] in der die Terminologie und die Regeln aus zwei Gründen auf dem Server hinterlegt werden: Erstens, damit sie zentral verwaltet werden; zweitens, damit sie nicht von den Anwendern (Clients) angepasst werden können (Drewer & Ziegler 2014: 232). Die Software wird in der Regel in die Autoren- oder Übersetzungsumgebung integriert (Drewer & Schmitz 2017: 201). Je nach Unternehmenssituation kann sie als Plug-in in dem Textbearbeitungsprogramm installiert werden, als ein Modul innerhalb eines großen Systems für Textproduktion und Übersetzung oder als ein alleinstehendes Programm (stand-alone) verwendet werden. Die CLCs können den Autor interaktiv darauf hinweisen, sobald ein Verstoß gegen eine der implementierten Regeln vorliegt; alternativ kann der Autor erst nach der Texterstellung die Prüfung starten (ebd.). Im Prüfergebnis werden die Verstöße hervorgehoben, kurz erläutert und dem Autor zusammen mit einem Alternativvorschlag angezeigt.

Im Redaktionsprozess führen die CLCs vier Funktionen durch: Prüfung der Rechtschreibung und Zeichensetzung, der Grammatik und des Stils sowie der korrekten Terminologieverwendung, wobei die wichtigsten Funktionen die Stilkontrolle und Terminologieprüfung darstellen (vgl. ebd.: 200).

Die grundlegende Funktionsweise eines CLCs lautet wie folgt (Siegel 2013: 52; Drewer & Ziegler 2014: 230): Vor der Identifizierung von Verstößen gegen die KS-Regeln führt das Programm zunächst primäre Schritte durch, wie Tokenisierung,[21] morphologische[22] und syntaktische[23] Analyse. Auf Basis der morphologischen Analyse werden vorwiegend die Rechtschreib-, Grammatik- sowie Terminologieprüfungen durchgeführt. Die Syntaktische Analyse ist hingegen für die Stilprüfung erforderlich.

[20] Eine Client-Server-Architektur ist „eine Systemarchitektur für verteilte Anwendungssysteme, bei der Subsysteme (Server) bestimmte Dienste anbieten, die von anderen Subsystemen (Clients) nutzbar sind". (Fettke 2016)

[21] Bei der Tokenisierung werden die einzelnen Wörter und Satzzeichen jedes Satzes identifiziert (Drewer & Ziegler 2014: 230).

[22] Bei einer morphologischen Analyse werden die Wörter in Morpheme zerlegt (Drewer & Ziegler 2014: 230).

[23] Bei einer syntaktischen Analyse werden zusammengehörige Gruppen von Satzelementen identifiziert (Drewer & Ziegler 2014: 230).

Nach Abschluss dieser Basisanalysen startet die Software die vier Prüfungen (Drewer & Ziegler 2014: 231): Die *Rechtschreibprüfung* ist umfangreicher als die bekannten Rechtschreibprüfungen von Textverarbeitungsprogrammen. Dank einer semantischen Komponente können Tippfehler, die zwar Wörter ergeben, aber semantisch im Satz falsch sind (z. B. „ins" und „uns"), erkannt werden. Ebenfalls wird die Zeichensetzung geprüft. Die *Grammatikprüfung* ist für die korrekte Grammatik zuständig. (Drewer & Schmitz 2017: 201)

Die *Stilprüfung* hat die Aufgabe, schwer verständliche und für den Texttyp nicht adäquate Konstruktionen zu identifizieren (Siegel 2013: 52; Drewer & Ziegler 2014: 231), wodurch die Verständlichkeit und somit die Qualität des Texts erhöht wird (Drewer & Schmitz 2017: 201). Beispiele für Stilregeln aus der tekom (2013) für deutsche technische Dokumentationen sind: „Konjunktiv zu vermeiden", „Anweisungen als imperativischen Infinitiv formulieren" und „Direkte Anrede verwenden". Zudem entwickelt das Unternehmen durch die Stilprüfung seine Corporate Language, denn die Prüfung erkennt die Sätze, die zwar grammatisch korrekt sind, aber nicht nach dem vorgegebenen Stil verfasst wurden (Drewer & Ziegler 2014: 232).

Für die *Terminologieprüfung* benötigt das Programm eine Terminologiedatenbank, die die Vorzugsbenennungen, unzulässige Benennungen und zulässige Benennungen enthält. Auf Basis der durchgeführten Sprachanalyse zusammen mit dieser Terminologiedatenbank kann das Programm unzulässige Flexionsvarianten identifizieren und dem Autor eine Vorzugbenennung vorschlagen. Eine weitere Aufgabe der Terminologieprüfung ist die Identifizierung von Wörtern, deren Schreibweise nicht regelkonform ist (Beispiel aus der tekom-Leitlinie „Komposita aus zwei Basismorphemen immer ohne Bindestrich"). (Siegel 2013: 52; Drewer & Ziegler 2014: 232) Ferner sind die CLCs mit einer Terminologieextraktionskomponente ausgerüstet. Durch die *Terminologieextraktion* werden domänenspezifische Termini automatisch erkannt und in einer vorgesehenen Tabelle gespeichert. Die Terminologieextraktion wird mithilfe von Regeln ausgeführt, die auf Part-of-Speech-Informationen und Lemmatisierung aufgebaut sind. (Siegel 2013: 52)

Die am meisten implementierten CLCs auf dem deutschen Markt sind die Softwareprodukte der Firmen „Acrolinx GmbH" und „Congree Language Technologies GmbH" (vgl. Geldbach 2009; Drewer & Ziegler 2014: 230). Die CLC-Komponente der Congree Software ist der Controlled-Language-Checker CLAT,[24] der vom Institut der Gesellschaft zur Förderung der Angewandten Informations-

[24] http://www.iai-sb.de/de/produkte/clat [abgerufen am 23.12.2014]

forschung (IAI)[25] der Universität des Saarlandes entwickelt wurde. Dank einer Forschungskooperation mit dem IAI wurde der vorliegenden Studie eine Lizenz für CLAT zu Forschungszwecken zur Verfügung gestellt. CLAT wurde in der Studie zur Prüfung der Verstöße gegen die analysierten KS-Regeln verwendet (siehe §4.5.3.1, Schritt [2]). Im nachstehenden Abschnitt erfolgt eine detaillierte Darstellung von CLAT.

2.6.3 CLAT – Controlled Language Authoring Technology

CLAT steht für Controlled Language Authoring Technology (Haller & Schütz 2001). Ziel der Software ist es, technische Redakteure bei der Erstellung hochwertiger Dokumentationen, z. B. nach bestimmten Standards, zu unterstützen (Rösener 2010). Die ersten deutschen Versionen von CLAT wurden bei BMW München und Heidelberger Druckmaschinen; und die ersten englischen Versionen bei Sun Microsystems implementiert (Haller & Schütz 2001). Die Software bietet keine Kontrollierte Sprache sondern eine Regelsammlung an, die auf langjähriger Forschung und fundierter Erfahrung des IAIs (Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung) basiert (Geldbach 2009). Das IAI hat bei der Entwicklung des „Technischen Deutsch“ von tekom mitgewirkt, daher sind die tekom-Regeln in CLAT abgedeckt (ebd.). Das erwerbende Unternehmen hat die Möglichkeit, vorhandene Regeln zu aktivieren bzw. deaktivieren (ebd.).

CLAT hat eine Client-Server-Architektur und deckt in seiner Prüfung die Bereiche Rechtschreibung, Grammatik, Stil, Terminologie und Termextraktion ab (Geldbach 2009; Rösener 2010). Anhand des Werkzeugs „UMMT“ (Utility for Mandate Management Tasks) erfolgt die Konfiguration der Regelwerke, Terminologie, Synonyme sowie Benutzerhandbücher entsprechend den Unternehmensanforderungen (Geldbach 2009). Konkret kann man in UMMT folgende Haupteinstellungen vornehmen: Terminologie importieren und verwalten, stilistische und grammatische Regeln aktivieren und deaktivieren sowie spezielle Schreibweisen und Synonyme definieren (Rösener 2010).

In einer primären Analyse wird der Text auf Grundlage der morphologischen und syntaktischen Analyse in Sätze und Wörter zerlegt (Haller & Schütz 2001). Bei der *Rechtschreibprüfung* werden falsche bzw. unbekannte Wörter markiert

[25] Im Jahr 2002 hat das IAI die Software CLAT als ein Upgrade von MULTILINT freigegeben (Ramírez Polo & Haller 2005). Bei den Bestandskunden ist die Software unter dem Namen CLAT im Einsatz. Seit 2011 wird CLAT nicht mehr von dem IAI sondern von der Congree Language Technologies GmbH vertrieben, an der das IAI beteiligt ist. (Drewer & Ziegler 2014: 230)

(Rösener 2010). Außerdem kann die Software im Rahmen der Rechtschreibprüfung zwischen Sprachvarianten (e.g. Britisches vs. Amerikanisches Englisch) unterscheiden (ebd.). Anders als die klassische Rechtschreibprüfung, die auf Basis eines Lexikons durchgeführt wird, erfolgt die Rechtschreibprüfung in CLAT auf Basis einer vollständigen Liste der Morpheme zusammen mit der unternehmensspezifischen Terminologie; so wird jedes Wort, das nach den morphologischen Regeln gebildet wird bzw. ein unternehmensspezifischer Terminus ist, als fehlerfrei erkannt (Haller & Schütz 2001). Auf diese Weise ist die Anzahl der als fehlerhaft markierten Wörter wesentlich geringer als die einer herkömmlichen Rechtschreibprüfung. Nicht selten führt ein Rechtschreibfehler zu einem lexikalisch korrekten Wort; diese Art von Rechtschreibfehlern kann nur durch die syntaktische Analyse des Satzes erkannt werden. (ebd.) Die *Grammatikprüfung* ermittelt grammatische sowie typographische Fehler (Rösener 2010). Hierbei werden Partial Parsings berechnet und Tests für Wort- oder Musterfehler durchgeführt (Haller & Schütz 2001). Der Stil im Sinne von Verständlichkeit, Klarheit und stilistischer Eignung ist die Aufgabe der *Stilprüfung* (Rösener 2010). Auf Basis von der unternehmensspezifischen Schreibregeln, die das Unternehmen bei der Konfiguration vom CLAT im UMMT festlegt, werden komplexe, ambige bzw. stilistisch problematische Stellen hervorgehoben. Die letzte Prüfungskomponente in CLAT beschäftigt sich mit der *Terminologie*. CLAT ermittelt im Text nicht zugelassene bzw. veraltete Termini und schlägt dafür die entsprechenden bevorzugten Termini vor. Nach der Prüfung hat der technische Redakteur die Möglichkeit die ermittelten Regelverstöße zu überprüfen, den Text zu überarbeiten oder keine Überarbeitung durchzuführen. Neben den vier Prüfungen bietet CLAT die *Termextraktionsfunktion*, mit der die Software den Terminologie-Workflow im Unternehmen unterstützt. Mithilfe dieser Funktion werden Substantive mit Terminieigenschaften ermittelt, so dass der technische Redakteur während der Prüfung entscheiden kann, ob sie als zulässig bzw. unzulässig in der Unternehmensdatenbank erfasst werden sollen. (ebd.)

Während der Anpassungsphase (customization phase) können die Fehlermeldungen und die darin enthaltenen Beispiele je nach Abteilung oder sogar Autor angepasst werden (Haller & Schütz 2001). Zudem können je nach Dokumenttyp unterschiedliche Regeln gelten; beispielsweise kann die Prüfung von Anleitungstexten strenger als die der Informationstexte gestaltet werden (ebd.). Schließlich unterstützen CL-Cherker das Unternehmen dabei, die technischen Dokumente in Bezug auf Lesbarkeit und Verständlichkeit zu verbessern. Dies wiederum stellt eine solide Grundlage für nachfolgende Prozesse wie die Übersetzung und Qualitätssicherung dar.

2.7 Fazit

Angesichts des zahlreichen Nutzens der Kontrollierten Sprache ist ihr Einsatz sehr weit verbreitet. Je nach den Unternehmensbedürfnissen wird sie in unterschiedlichem Umfang – angefangen bei einzelnen Regeln bis hin zu vollständiger KS – eingesetzt. Für die deutsche Sprache wird meist nach einem Regelwerk mit grammatischen und syntaktischen Regeln, terminologischen Vorgaben sowie Fachtermini (ohne den Einsatz eines Lexikons) gearbeitet. Mithilfe der KS zielen die Unternehmen auf eine bessere Lesbarkeit, Verständlichkeit und Übersetzbarkeit ab. Je nach Dokumenttyp können die zweckdienlichen KS-Regeln mithilfe von CL-Checkern aktiviert werden. Die Aktivierung der Regeln bewirkt im Allgemeinen eine vereinfachte Satzstruktur und -komplexität bzw. eine reduzierte Ambiguität. Das wiederum wirkt sich auf die Verständlichkeit, Lesbarkeit sowie Übersetzbarkeit – wenn auch in unterschiedlichem Ausmaß – aus. Daher sind die KS-Auswirkungen voneinander nicht scharf zu trennen. Vor diesem Hintergrund und angesichts des aktuell bemerkbaren MÜ-Fortschritts insbesondere bei der Lieferung eines flüssigen und stilistischen Outputs nimmt die Studie diverse KS-Regeln unter die Lupe, und zwar Regeln, die die Verständlichkeit und die Lesbarkeit, sowie weitere, die die Übersetzbarkeit, im Fokus haben, und analysiert ihre direkten bzw. indirekten Auswirkungen auf den MÜ-Output sowohl stilistisch als auch inhaltlich im Sinne der Verständlichkeit und Genauigkeit. Um all diese Qualitätsaspekte abdecken zu können, wird die Analyse auf Basis humaner sowie automatischer Evaluationen durchgeführt. Da die maschinelle Übersetzung (MÜ) einer der Hauptakteure der vorliegenden Arbeit ist, wird das Augenmerk im folgenden Kapitel auf die MÜ sowie vorherige MÜ-Studien im Kontext der KS gerichtet.

3 Maschinelle Übersetzung

3.1 Einleitung

Mit der stetigen Zunahme der zu übersetzenden Datenmenge stieg auch der Bedarf an maschinellen Übersetzungen (MÜ). Viele Unternehmen integrieren heutzutage die MÜ in ihren Übersetzungsprozess und versuchen mit der Einbettung eines anschließenden Post-Editing-Schritts in ihren Workflow, die Übersetzungsproduktivität zu erhöhen und gleichzeitig die Übersetzungskosten zu reduzieren. Dieser Bedarf gekoppelt mit steigenden Qualitätsansprüchen führt dazu, dass MÜ-Systeme ständig unter die Lupe genommen werden. Das vorliegende Kapitel beschäftigt sich mit der Entwicklung, den Ansätzen und der Qualitätsbewertung der MÜ. Auf die Definition der MÜ und eine Erläuterung ihrer Relation zu der Kontrollierten Sprache (KS) folgt eine Darstellung der verschiedenen MÜ-Ansätze. Anschließend befasst sich das Kapitel mit der Thematik der MÜ-Qualität und den verschiedenen Qualitätsbewertungstechniken. Zum Schluss werden MÜ-Studien im Kontext der KS diskutiert; auch die geringe Anzahl der KS-Untersuchungen auf Regelebene und die Herausforderungen solcher Untersuchungen werden näher betrachtet.

3.2 MÜ – Begriffsbestimmung und Motiv des KS-Einsatzes

Die maschinelle Übersetzung wird als „der Prozess der automatischen Übersetzung von Text aus einer natürlichen Sprache in eine andere“ definiert (Dorr u. a. 1999). Inwiefern MÜ-Systeme selbständig bzw. ohne menschliche Intervention funktionieren können, ist ein wesentlicher Punkt von vielen Definitionen der MÜ. Im Jahr 1987 definierte Goshawke die MÜ als

> the transfer of meaning from one natural (human) language to another with the aid of a computer. There are very few systems that are, or even attempt to be, complete machine translation systems in themselves – nearly all systems are Machine Aided Translation (MAT), involving human help either at the input stage (pre-editing) or the output stage (post-editing) or both. (Goshawke u. a. 1987: 6)

Auch bei (Hutchins & Somers 1992: 3) ist das menschliche Eingreifen ein Bestandteil der Definition der MÜ-Systeme, so seien „computerized systems responsible for the production of translations from one natural language into another, with or without human assistance". Ein Verzicht auf die menschliche Intervention in der MÜ ist bisher als unrealistisch zu betrachten. In den 50er Jahren kritisierte Yehoshua Bar-Hillel die Annahme, dass eine Fully Automatic High Quality Translation (FAHQT) das Ziel der MÜ sei (ebd.: 6f.). Nach wie vor bezeichnet man den MÜ-Output als Rohübersetzung und versucht mithilfe des Pre-Editing bzw. des Post-Editing diesen Rohoutput zu verbessern und die Übersetzungsproduktivität zu erhöhen.

Gleichzeitig erlebte die MÜ-Entwicklung große technologische Sprünge von einem Ansatz zum anderen – angefangen bei den regelbasierten Systemen über die statistischen und hybriden Systeme bis hin zu den neuronalen Systemen. Mit dieser Entwicklung verbesserte sich der MÜ-Output in den verschiedenen Domänen zu einem unterschiedlichen Grad. Der Qualitätsunterschied des MÜ-Outputs ist leicht erkennbar, wenn man in einem Extremfall einen Abschnitt aus einem Benutzerhandbuch und ein kurzes Gedicht maschinell übersetzt. Die MÜ-Qualität des technischen Texts fällt in der Regel viel höher als die des Literarischen aus. Wenn ein MÜ-System theoretisch in der Lage ist, eine hohe Qualität zu liefern, warum ist dies nicht immer realisierbar? Hutchins & Somers (1992: 2) geben eine schlüssige Antwort auf diese Frage: „The major obstacles to translating by computer are, as they have always been, not computational but linguistic." Je einfacher der Text auf linguistischer Ebene ist, desto höher fällt die Qualität seiner MÜ aus. Dementsprechend wurde durch die Anwendung von Pre-Editing-Techniken wie der Kontrollierten Sprache daran gearbeitet, durch die Vermeidung von Ambiguität, Vereinfachung der Satzstruktur und Einschränkung bzw. Standardisierung der Lexik, die Komplexität des Ausgangstexts zu reduzieren. Bereits in der ersten Konferenz für MÜ 1952 umfassten die Pläne und Anregungen für zukünftige Forschung das Schreiben in Kontrollierten Sprachen, die Konstruktion von Subsystemen sowie die Anerkennung des Bedarfs an menschlicher Unterstützung in Form von Pre- und Post-Editing (ebd.: 6).

In der vorliegenden Arbeit liegt der Fokus auf dem Einsatz der Kontrollierten Sprache (KS) und ihrem Einfluss auf den MÜ-Output. Durch einen Vergleich der unterschiedlichen MÜ-Ansätze wird in der Arbeit untersucht, ob und inwiefern die analysierten Regeln der KS heutzutage die MÜ-Qualität verbessern können und somit den menschlichen Eingriff in den Übersetzungsprozess reduzieren.

3.3 Entwicklung der MÜ-Ansätze

Die Forschung in diesem Bereich begann nach dem Zweiten Weltkrieg, zunächst mit dem Ziel, russische Texte ins Englische zu „decodieren“. Die Begriffe, die damals aus der Kryptographie entlehnt wurden, werden bis heute weiterhin verwendet. Leon Dostert arbeitete 1954 an der Georgetown Universität gemeinsam mit IBM an einem Projekt, im Rahmen dessen sorgfältig ausgewählte russische Beispielsätze mit einem sehr begrenzten Vokabular von 250 Wörtern und sechs Grammatikregeln ins Englische übersetzt wurden (Hutchins & Somers 1992: 6). Dem lag die Idee zugrunde, auf diese Weise die technische Durchführbarkeit der MÜ nachzuweisen. Die ersten entwickelten Systeme waren regelbasiert, gefolgt von den SMÜ-Systemen. Um von den Vorteilen der beiden Ansätze zu profitieren, wurden hybride Systeme entwickelt, die auf regelbasierten und statistischen Komponenten beruhen. Der aktuellste MÜ-Ansatz ist der neuronale Ansatz, der von dem massiven Volumen der gesammelten Daten profitiert und auf Basis von neuronalen Netzen arbeitet. In dieser Studie wurden MÜ-Systeme verwendet, die nach den folgenden Ansätzen konzipiert sind: ein regelbasiertes MÜ-System (RBMÜ), ein statistisches MÜ-System (SMÜ), zwei hybride MÜ-Systeme (HMÜ) und ein neuronales MÜ-System (NMÜ).[1] Im Folgenden werden die Entwicklung dieser Ansätze – chronologisch – dargestellt sowie deren zugrundeliegenden Paradigmen genauer erläutert.

3.3.1 Regelbasierte MÜ-Systeme

Hauptmerkmal der RBMÜ ist, dass dabei das System auf linguistische Regeln zurückgreift (Ramlow 2008: 38). Das System analysiert den Text auf verschiedenen linguistischen Ebenen, z. B. die Kasusflexionen auf morphologischer Ebene und Wortstellung auf syntaktischer Ebene. Danach – je nach RBMÜ-Methode – überführt das System den Ausgangstext in eine abstrakte Repräsentation. Anschließend wendet das System zur Produktion des Zieltexts Übersetzungsregeln an. (Werthmann & Witt 2014: 87)

Die ersten MÜ-Systeme waren regelbasiert. In den 60er Jahren entwickelte die Saarbrücken-Gruppe ein mehrsprachiges „Transfersystem“ namens SUSY. 1976 wurde Systran als RBMÜ-System von der EU-Kommission für die Übersetzung im Sprachenpaar Englisch-Französisch installiert und in den darauffolgenden Jahren um andere Sprachenpaare erweitert. In den 80er Jahren wurden ebenfalls die ersten RBMÜ-Systeme für die Sprachenkombination Englisch-Japanisch entwickelt. (Hutchins & Somers 1992: 7f.)

[1] Die Auswahlkriterien der untersuchten MÜ-Systeme sind unter §4.5.1 aufgeführt.

Die Architektur eines RBMÜ-Systems lässt sich mithilfe des Dreiecks von Vauquois anschaulich darstellen (Vauquois 1968):

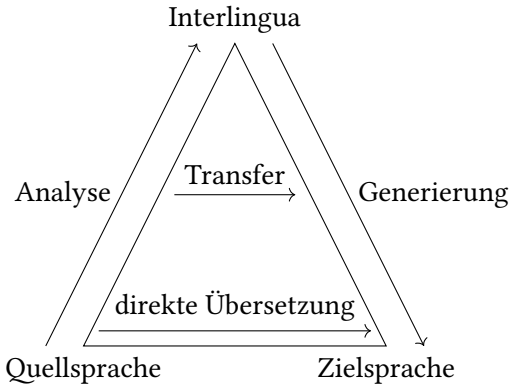


Abbildung 3.1: Vauquois-Dreieck der MÜ-Systemparadigmen. Quelle: Hutchins & Somers 1992: 107

In diesem Schema steigt die Systemkomplexität, je mehr man sich im Dreieck nach oben bewegt, von bloßen Wörtern über syntaktische Übertragungsregeln bis hin zu einer vollständig sprachunabhängigen Repräsentation (Interlingua). Wie das Dreieck von Vauquois zeigt, entwickelte sich der RBMÜ-Ansatz im Laufe der Zeit von der direkten Methode zu der transferbasierten Methode und anschließend zur Interlingua-Methode. Die Unterscheidungsmerkmale bei den Methoden sind die Anzahl der Sprachenpaare, die Analysemethode der aus dem Quelltext gewonnenen Informationen sowie die Methode der abstrakten Repräsentation dieser Informationen für die Übersetzung. (Werthmann & Witt 2014: 87) Im Folgenden eine kurze Erläuterung der drei Methoden:

Direkte Übersetzung: Bei dieser Methode werden die Quellsprachwörter wortweise mithilfe eines zweisprachigen Wörterbuchs in die Zielsprachwörter übersetzt. Die Wortstellung des Outputs erfolgt unter Verwendung einfacher syntaktischer Regeln (Stein 2009). Somit basiert diese Methode auf einer morphologischen Analyse der Wörter und einem komplexen zweisprachigen Wörterbuch sowie einigen einfachen Regeln zur Wortstellung (Carstensen u. a. 2004: 566). Es erfolgt keine Tiefenanalyse des Ausgangssatzes und es werden keine komplexen Regeln zur Generierung des Zielsatzes angewendet. Entsprechend ist die Qualität der Übersetzung im Allgemeinen nicht als hoch einzustufen und steigt, je ähnlicher die Ausgangs- und Zielsprachen einander syntaktisch und semantisch sind (Werthmann & Witt 2014: 88).

Transferbasierte Methode: In der transferbasierten Methode findet eine komplexe linguistische Analyse statt und beginnt die Generierung während der Übersetzung. Im Rahmen dieser Methode erfolgt die Übersetzung auf Satz- und nicht auf Wortebene (Werthmann & Witt 2014: 89). Die Methode besteht aus drei Schritten (Carstensen u. a. 2010: 646): Analyse, Transfer und Generierung. Im ersten Schritt wird der Quelltext geparst, segmentiert und syntaktisch bzw. semantisch analysiert. Im zweiten Schritt wird ein Transfer von der syntaktischen/semantischen Repräsentation des Ausgangssatzes in die Repräsentation der Zielsprache unter Verwendung von Mapping-Regeln durchgeführt. Zum Schluss wird der Zielsatz auf Basis der transferierten Repräsentation generiert. (Werthmann & Witt 2014: 89) Durch die syntaktische und semantische Sprachanalyse ist die Qualität der Übersetzung in der Transfermethode höher als die der direkten Methode (Stein 2009).

Interlingua-Methode: Im Gegensatz zu der transferbasierten Methode, die ausgangs- und zielsprach abhängige Regeln für den lexikalischen, syntaktischen und semantischen Transfer beinhaltet, arbeitet die Interlingua-Methode mit einer generischen Repräsentation, der sogenannten Interlingua (Werthmann & Witt 2014: 90). Die Grundidee der Interlingua besteht darin, den Ausgangssatz in einer sprachunabhängigen, abstrakten Begriffsrepräsentation darzustellen, die aus jeder Ausgangssprache erzeugt werden kann und zur Generierung der Übersetzung in eine beliebige Zielsprache verwendet wird (ebd.). Wie das Vauquois-Dreieck zeigt, ist diese Methode im Vergleich zu den anderen Übersetzungsmethoden mit deutlich mehr Aufwand verbunden. Zusammenfassend besteht die Interlingua-Methode aus zwei Schritten (Carstensen u. a. 2010: 646): Analyse und Generierung. Im Analyseschritt wird der Ausgangstext geparst, segmentiert und analysiert, um die Interlingua der Ausgangssätze zu erzeugen. Im Generierungsschritt wird der Zieltext aus der Interlingua erzeugt. Entsprechend wird der Vorteil dieser Methode als eng an ihren Nachteil gebunden betrachtet (Werthmann & Witt 2014: 90f.): Einerseits erspart die ausgangsprachunabhängige Zwischen-repräsentation die Erzeugung von sprachspezifischen lexikalischen, syntaktischen und semantischen Regeln bei jedem Sprachenpaar; andererseits stellt die Realisierung solcher generischen Zwischenrepräsentationen für eine formale Sprache mit eigenen lexikalischen und syntaktischen Elementen eine komplexe Aufgabe dar. Daher wird diese Methode für bestimmte Domänen (z. B. Hotelreservierung) oder für kontrollierten Zielsprachen mit eingeschränktem Vokabular und begrenzten Satzstrukturen verwendet (Al-Ansary 2011).

Die Ergebnisqualität der RBMÜ-Systeme variiert von einem Sprachenpaar zum anderen und hängt davon ab, ob eine bestimmte Fachsprache unterstützt wird und die notwendigen Fachtermini eingepflegt sind (Stein 2009). Daher stellt die

aufwendige Bereitstellung von zweisprachigen Wörterbüchern und Regelwerken einen Nachteil von RBMÜ-Systemen dar (Werthmann & Witt 2014: 91). Dies ließ die statistischen MÜ-Systeme seit Ende der 80er Jahre als präzisere und weniger aufwendige Systeme in den Vordergrund rücken (Tripathi & Sarkhel 2010).

Im Rahmen dieser Studie wird der KS-Effekt auf den RBMÜ-Output des Systems *Lucy LT* [2] untersucht. Lucy LT (Nachfolger des METAL MÜ-Systems) ist ein transferbasiertes MÜ-System, das mit einem Island-Chart-Parser und in drei Übersetzungsphasen arbeitet (Martin & Serra 2014): Analyse, Transfer und Generierung. In jeder dieser Phasen und für jede Sprachrichtung werden eine Analysegrammatik, eine Transfergrammatik und eine Generierungsgrammatik sowie Ausgangs- und Zielsprachenlexika und ein Ausgang-zu-Ziel-Transferlexikon verwendet (ebd.).

3.3.2 Statistische MÜ-Systeme

Anstelle der Bereitstellung von aufwendigen Regelsammlungen zur Erstellung von RBMÜ-Systemen überraschte der IBM-Wissenschaftler Peter Brown 1988 das Publikum auf der zweiten TMI-Konferenz mit dem Konzept der Statistischen MÜ (SMÜ), das auf parallelen Korpora basiert (Stein 2009). Ein reines SMÜ-System verwendet keine linguistischen Informationen (Forcada 2010). Die Grundidee eines SMÜ-Systems besteht darin, dass „a source language and a target language sentence are a translation of each other with a certain probability" (ebd.: 220). Somit ist das Ziel eines SMÜ-Systems, den entsprechenden Satz in der Zielsprache zu finden, der die höchste Wahrscheinlichkeit aufweist (ebd.).

In der Klassifizierung der MÜ-Ansätze fällt die SMÜ zusammen mit der beispielbasierten MÜ in die Kategorie der korpusbasierten Methoden. Hauptmerkmal dieser Methoden ist die Übersetzung mithilfe eines zweisprachigen Korpus. Der zu übersetzende Satz wird mit dem Korpus abgeglichen, um eine Übersetzungslösung zu finden. In der *beispielbasierten MÜ* greift das System auf Beispielübersetzungen in einem zweisprachigen Korpus zurück und ordnet dem Ausgangssegment das am ehesten vergleichbare Zielsegment zu. Bei der *SMÜ* wird das zweisprachige Korpus, das als Übersetzungsmodell agiert, durch einsprachige Korpora für die beteiligten Sprachen, die als Sprachmodelle fungieren, ergänzt. (Ramlow 2008: 42f.)

Abbildung 3.2 veranschaulicht die Architektur eines typischen SMÜ-Systems, das aus den folgenden Komponenten besteht (Petrzelka 2011: 8ff.; Dubey 2017):

[2] Link zum System Lucy LT: http://www.lucysoftware.com/english/machine-translation/lucy-lt-kwik-translator-

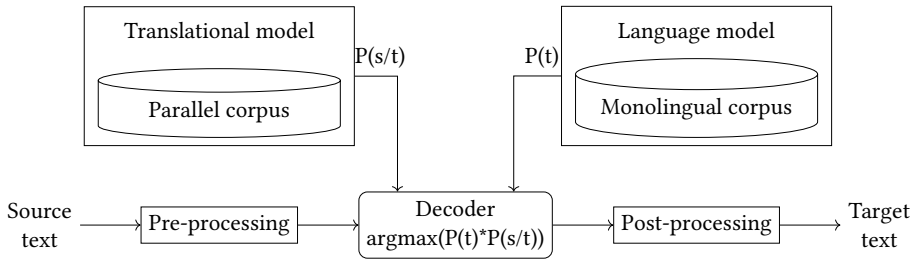


Abbildung 3.2: SMÜ-Architektur. Quelle: Dubey 2017

(1) einem *Zielsprachenmodell*, das auf einem einsprachigen Korpus in der Zielsprache basiert und für die korrekte Zusammenstellung der Wörter in der Zielsprache zuständig ist. Es berechnet die Wahrscheinlichkeit des Zielsatzes;

(2) einem *Übersetzungsmodell*, das auf einem Parallelkorpus basiert und die bedingte Wahrscheinlichkeit des Zielsatzes in Anbetracht des Ausgangssatzes berechnet. Um dies zu realisieren, erfolgt zunächst ein Word-Alignment, bei dem ermittelt wird, welches Wort im Ausgangssatz welchem Wort im Zielsatz entspricht. Auf Basis des Word-Alignments wird ein Phrasen-Alignment durchgeführt und die sogenannte „Phrasenübersetzungstabelle" (phrase-translation table) erstellt, d. h. eine Tabelle, die die Phrasen des Ausgangstexts, die Phrasen des Zieltexts sowie deren Übereinstimmungswahrscheinlichkeit (matching probability) beinhaltet;

(3) einem *Decoder*, der die eigentliche Übersetzungsaufgabe übernimmt, indem er das bestmögliche Übersetzungspaar durch das Maximieren der zwei zuvor genannten Wahrscheinlichkeiten findet. Je häufiger ein Übersetzungspaar in den Korpora auftritt, desto höher ist sein Ranking.

Man unterscheidet im SMÜ-Ansatz zwischen der wortbasierten und phrasenbasierten SMÜ, je nachdem auf welcher Ebene das Übersetzungsmodell erstellt wird und ob die Wahrscheinlichkeiten für Wörter oder Phrasen berechnet werden (Werthmann & Witt 2014: 93):

Wortbasierte SMÜ: Hierbei erfolgt die Analyse der Korpora auf Wortebene. Der Satz wird in Wörter aufgeteilt und auf Wortebene übersetzt (Köhn 2010: 81). Dementsprechend setzt dieser Ansatz eine 1:1-Beziehung zwischen den Wörtern in der Ausgangs- und Zielsprache voraus, was in der Realität jedoch nicht gegeben ist (Werthmann & Witt 2014: 94). Daraus entstand die Notwendigkeit der Übersetzung auf Phrasenebene.

Phrasenbasierte SMÜ (PBMÜ): Bei diesem Ansatz stellt eine Phrase eine Übersetzungseinheit dar, wobei es sich hier nicht um eine Phrase im linguistischen Sinn handelt, sondern um eine syntaktische Phrase, die aus einer Sequenz von Wörtern besteht (Köhn 2010: 128). Sobald die Phrasen in die Zielsprache übersetzt wurden, übernimmt das Sprachmodell seine Aufgabe, nämlich die Phrasen in der richtigen zielsprachigen Reihenfolge zu ordnen (Carstensen u. a. 2010: 650). Ein wichtiger Vorteil der phrasenbasierten gegenüber der wortbasierten SMÜ beruht auf der Möglichkeit, ein Wort des Ausgangssatzes mit mehreren Wörtern in den Zielsatz zu übersetzen und umgekehrt (Werthmann & Witt 2014: 95). Ein sehr bekanntes phrasenbasiertes SMÜ-System ist Moses. In Moses wurde die phrasenbasierte Übersetzung (PBMÜ) um Faktoren und Confusion-Network-Decoding erweitert (Koehn u. a. 2007): Ein Factored-Übersetzungsmodell erlaubt die Verwendung morphologischer, syntaktischer und semantischer Informationen. Die Integration des Confusion-Network-Decoding ermöglicht die Übersetzung von mehrdeutigen Inputs, da – anstatt der bloßen Ermittlung des besten Outputs – ein Netzwerk von unterschiedlichen Wortoptionen untersucht wird.

Im Vergleich zu den RBMÜ-Systemen sind die SMÜ-Systeme günstiger (Werthmann & Witt 2014: 95), da die zeitaufwendige Erstellung von Regelwerken entfällt. Wenn Parallelkorpora für weitere Sprachen verfügbar sind, lassen sich die SMÜ-Systeme ohne großen Aufwand um weitere Sprachen erweitern (Stein 2009), jedoch ist die Erstellung von großen alignierten Korpora mit Aufwand verbunden, was als Schwäche der SMÜ betrachtet wird (Werthmann & Witt 2014: 96).

Im Rahmen dieser Studie wird das SMÜ-System *SDL Free Translation*[3] untersucht. SDL Free Translation ist ein rein statistisches MÜ-System der Firma SDL.[4]

3.3.3 Hybride MÜ-Systeme

Nachdem die RBMÜ als der traditionelle Ansatz der MÜ im Laufe der 90er Jahre durch den SMÜ-Ansatz zunehmend ersetzt wurde, kam die Idee auf, beide Ansätze zu kombinieren, um gleichzeitig von ihren Vorteilen zu profitieren und ihre Schwächen gegenseitig kompensieren zu können. In der hybriden Methode liegt der Fokus auf der Kombinierung der besten Eigenschaften von zwei oder mehr MÜ-Ansätzen, häufig werden dabei Regeln in einen SMÜ-Ansatz integriert (Costa-Jussà & Fonollosa 2015).

[3] Link zum System SDL Free Translation: https://www.freetranslation.com/de

[4] Quelle: „What Is Machine Translation and How Does It Work?“, Online unter: https://sdl.uservoice.com/knowledgebase/articles/256030-what-is-machine-translation-and-how-does-it-work [abgerufen am 06.12.2016].

Entsprechend kam zur Jahrtausendwende der Trend auf, HMÜ-Systeme zu entwickeln. Cavalli-Sforza & Lavie (2006) beschreiben die Verbreitung von hybriden Systemen sowohl in der Forschung als in der Industrie wie folgt:

> Much current research in machine translation is neither based purely on linguistic knowledge nor on statistics, but includes some degree of hybridization. At AMTA 2004 and MT Summit 2005 just about all commercial MT developers also claimed to have hybrid systems. (Cavalli-Sforza & Lavie 2006: 1)

Die hybriden Systeme werden unterschiedlich aufgebaut; sie beinhalten verschiedene RBMÜ- und SMÜ-Komponenten. Das System kann z. B. Module enthalten, in denen die Wörterbücher und Regeln des Systems sich je nach Übersetzungsdomäne (IT, Technik, Medizin) dynamisch anpassen. Eine weitere Aufbaumöglichkeit ist der Einsatz eines statistischen Ansatzes zur Unterstützung bei einer korrekten Übersetzung auf Wortebene (z. B. das Wort „Maus“ in der IT- vs. Zoologie-Domäne). (Poibeau 2017: 171)

Aufgrund der Diversität der Aufbaumöglichkeiten existieren mehrere Klassifizierungen für die HMÜ-Systeme. Eisele (2007) unterscheidet bei der Hybridisation zwischen einer flachen Integration, in der zwei oder mehr Systeme in einem größeren System integriert werden, und einer tiefen Integration, in der die Vorteile von zwei Ansätzen zusammengeführt werden. Eine andere Klassifizierung von Way (2010) unterscheidet zwischen einem Multi-Engine-Ansatz und einem integrierten Systemansatz. Dennoch, wie oben erwähnt, sind die meisten hybriden Systeme eine Kombination von RBMÜ- und SMÜ-Systemen. Daher wäre es an dieser Stelle sinnvoll zwischen zwei Varianten der Hybridisation zu unterscheiden (Costa-Jussà & Fonollosa 2015): (1) ein RBMÜ-System, das mit Data ergänzt wurde (eine RBMÜ-geführte Hybridisation) und (2) ein SMÜ-System, das mit linguistischen Regeln ergänzt wurde (eine SMÜ-geführte Hybridisation).

RBMÜ-geführte Hybridisation: Diese Art von Hybridisation kann auf mehrere Weise realisiert werden, z. B. durch die Einführung eines Korpus zum Aufbau des RBMÜ-Systems, die Einführung von korpusbasierten Tools zur Gewichtung des RBMÜ-Outputs oder die Durchführung eines statistischen Post-Editing für den RBMÜ-Output (ebd.).

Die RBMÜ wurde kritisiert, dass der RBMÜ-Output im Vergleich zum SMÜ-Output oft weniger natürlich und flüssig klingt, da RBMÜ-Systeme keinen Zugriff auf statistische Sprachmodelle haben. Diese Schwäche wird im hybriden Ansatz durch die Integration eines Flüssigkeitsmodells in die RBMÜ-Architektur

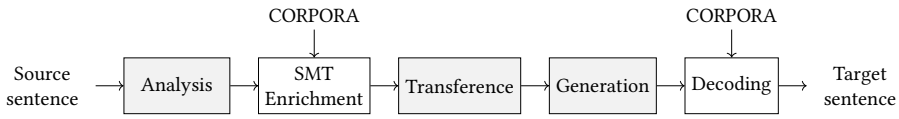


Abbildung 3.3: RBMÜ-geführte Hybridisation. Quelle: Costa-Jussà & Fonollosa 2015

mittels Post-Editing behandelt. Auf diese Weise konnte der MÜ-Output eines reinen RBMÜ-Systems im Hinblick auf die Flüssigkeit und die Kontextualität verbessert werden. (Eisele u. a. 2008)

SMÜ-geführte Hybridisation: Diese Art von Hybridisation erfolgt durch die Kombinierung verschiedener korpusbasierter MÜ-Ansätze oder durch die Integration von Regeln in ein korpusbasiertes MÜ-System, und zwar mittels der Verwendung von Regeln bei dem Pre-/Post-Editing oder der Integration von Wörterbüchern bzw. Regeln in das Kernmodell (Costa-Jussà & Fonollosa 2015).

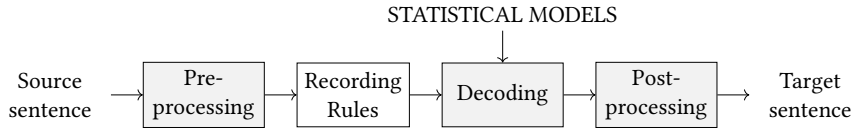


Abbildung 3.4: SMÜ-geführte Hybridisation. Quelle: Costa-Jussà & Fonollosa 2015

Die Ergänzung von SMÜ-Systemen um syntaktische Analysen oder semantische Operationen kann insbesondere bei Sprachenpaaren, für die keine großen Korpora verfügbar sind oder deren Satzbau und Flexion sich stark unterscheiden, von Vorteil sein (Stein 2009).

Zusammenfassend wird in den verschiedenen Hybridarchitekturen in der Regel entweder eine RBMÜ-Engine verwendet, um die für den SMÜ-Decoder verfügbaren lexikalischen Ressourcen anzureichern; oder es werden Komponenten der SMÜ-Infrastruktur zusammen mit linguistischer Verarbeitung und manueller Validierung verwendet, um das Lexikon einer RBMÜ-Engine zu erweitern (vgl. Eisele u. a. 2008). Je nach Art der Kombination der Hybridstruktur, kann ein HMÜ-System in verschiedenen Szenarien unterschiedliche Stärken und Schwächen zeigen.

Aufgrund der Aufbaudiversität in diesem MÜ-Ansatz wurden in dieser Studie zwei verschiedene Hybridsysteme untersucht: *Systran* und *Bing Translator.*[5]

[5]Links zu den Systemen Systran und Bing Translator: http://www.systranet.com/translate, https://www.bing.com/translator

Bing Translator ist ein „statistisches MÜ-System mit sprachspezifischen Regelkomponenten für das Zerlegen und Zusammensetzen von Sätzen“ (Werthmann & Witt 2014: 84), das von Microsoft Übersetzer angetrieben wird. Systran wurde ursprünglich als RBMÜ entwickelt und erst 2009 durch die Ergänzung um statistische Komponenten zu einem Hybridsystem weiterentwickelt (ebd.: 99).

3.3.4 Neuronale MÜ-Systeme

Nach dem Fortschritt, den die SMÜ bis zum Ende des 20. Jahrhunderts verzeichnete, waren keine weiteren Durchbrüche zu beobachten. Die SMÜ erreichte eine Stufe, nach der eine markante Verbesserung schwer realisierbar war. So gaben die Forscher von Google Translate im Jahr 2010 im „Guardian“ zu, dass

> the idea that more and more data can be introduced to make the system better and better is probably a false premise. Each doubling of the amount of translated data input led to about 0.5% improvement in the quality of the output. […] So now it is much more important again to add on different approaches […]. (Adams 2010: 4)

Vor diesem Hintergrund entstand die Notwendigkeit, neue Ansätze zu entwickeln und gleichzeitig von dem massiven Volumen an gesammelten Daten zu profitieren. Zu dieser Zeit hat der Ansatz des *Machine Learning* auf Basis von neuronalen Netzen im Bereich der Bildverarbeitung signifikante Erfolge verzeichnet. Dies motivierte die MÜ-Forscher, damit zu beginnen, MÜ-Modelle nach Deep-Learning-Ansätzen zu entwickeln. Die NMÜ-Forschung wurde nicht nur in den meisten akademischen Zentren der Computational Linguistics auf der ganzen Welt durchgeführt, sondern auch in global agierenden Internetfirmen wie Facebook, Microsoft, Google, SDL und Systran. (Poibeau 2017: 45)

Die Kernidee des Deep Learning bei der Bilderkennung besteht darin, die relevantesten Merkmale (features) basierend auf der Verarbeitung von zahlreichen Beispielen abzuleiten. So werden – bei der Bilderkennung – komplexere Strukturen ausgehend von sehr vielen Pixeln erkannt. Analog dazu werden – bei Sprachen – Sequenzen von Wörtern oder Phrasen auf Basis von Zeichen oder Wörtern identifiziert. Das Konzept simuliert somit die menschliche Wahrnehmung, in der das Gehirn Gruppen von einfachen Items sehr schnell analysiert, um übergeordnete Merkmale zu identifizieren, oder von partiellen Informationen eine komplexe Repräsentation ableitet. (ebd.: 183)

Mit diesem Konzept versucht man, die weiterhin bestehenden Herausforderungen der MÜ zu bewältigen, wie z. B. die Herausforderung, dass der Sinn eines Worts je nach Kontext und Zielsprache auf verschiedene Weise übersetzt

bzw. mit einem Wort oder einer Wortgruppe ausgedrückt werden kann (ebd.: 182). Während die SMÜ kontextunabhängig auf Phrasenebene die Übersetzung produziert, lautet die Hypothese der NMÜ, dass Wörter, die in ähnlichen Kontexten vorkommen, eine ähnliche Bedeutung haben können. Entsprechend versucht die NMÜ, Wörter zu identifizieren und gruppieren, die in ähnlichen translatorischen Kontexten in sog. *Worteinbettungen* (word embeddings) auftreten. Auf diese Weise können die NMÜ-Systeme besser mit seltenen Wörtern sowie Wörtern mit verschiedenen Bedeutungen umgehen als frühere MÜ-Ansätze: Bei seltenen Wörtern können andere Wörter, die in ähnlichen Kontexten auftreten, auf eine nützliche Übersetzung hinweisen; Wörter, die verschiedene Bedeutungen haben, können zu verschiedenen Einbettungen gehören. (ebd.: 186)

Bei der Umsetzung des Konzepts des Deep Learning in der MÜ erstreben die NMÜ-Systeme das Lernen von komplexen Merkmalen völlig autonom und schrittweise aus den Daten, vollkommen ohne menschliche Anstrengung. Vor diesem Hintergrund besteht ein NMÜ-System aus einem Encoder und einem Decoder (siehe Abbildung 3.5).[6] Der Encoder hat die Aufgabe, die Trainingsdaten zu analysieren. Der Decoder ist dafür zuständig, die Übersetzung eines Satzes basierend auf den vom Encoder analysierten Daten automatisch zu erstellen. Anders als die SMÜ-Systeme arbeiten die Encoder und Decoder nicht auf Basis eines Sprachmodells und Übersetzungsmodells, sondern auf Basis eines neuronalen Netzwerks. In einem neuronalen Netzwerk wird jedes Wort durch einen Vektor von Zahlen kodiert, und alle Wortvektoren werden schrittweise kombiniert, um eine Repräsentation des gesamten Satzes zu liefern. (Xing u. a. 2016)

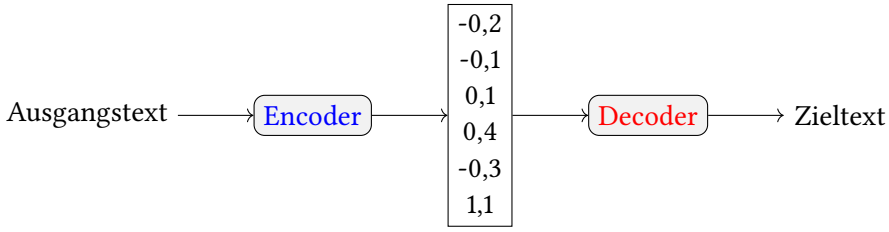


Abbildung 3.5: NMÜ-Architektur. Quelle: Fabienne 2016: 12

Die ersten Modelle der NMÜ gehen auf Ñeco & Forcada (1997) zurück. Die früheren NMÜ-Implementationen basierten auf einem konvolutionellen neuronalen Netzwerk (convolutional neural network) (Kalchbrenner & Blunsom 2013). Später wurden die Modelle durch die Verwendung wiederkehrender neuronaler Net-

[6] Zahlreiche Online-Tutorials erklären das NMÜ im Detail, zum Beispiel: http://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

ze (Recurrent Neural Networks, RNN) wesentlich verbessert (Cho u. a. 2014; Sutskever u. a. 2014). In einem RNN wird ein Quellsatz variabler Länge in einen Vektor fester Länge kodiert; dieser Vektor wird dann in einen Zielsatz mit variabler Länge dekodiert (Bahdanau u. a. 2015). Die Funktionsweise auf Basis eines End-to-End neuronalen Netzwerks ist die Hauptstärke der NMÜ (Wu u. a. 2016). Die Encoder-Decoder-Architektur ist in der Lage, den Quelltext in sog. Kontextvektoren, eine interne Repräsentation mit fester Länge, zu kodieren. Mit den Kontextvektoren können verschiedene Dekodiersysteme verwendet werden, um sie in verschiedene Sprachen zu übersetzen (Johnson u. a. 2016). Dennoch verursachte die Repräsentation mit fester Länge in der Encoder-Decoder-Architektur ein Problem bei der Übersetzung von langen Sequenzen (Pouget-Abadie u. a. 2014). Um dieses Problem anzugehen, wird ein Attention-Mechanismus verwendet. Durch den Attention-Mechanismus kann das Modell lernen, an welchen Stellen es auf die Quellsequenz fokussieren soll, um die semantischen Details zu erfassen, die zum Dekodieren jedes Wortes der Zielsequenz erforderlich sind. (Bahdanau u. a. 2015; Luong u. a. 2015) Darüber hinaus wurde das RNN-Modell kritisiert: Da dieses Modell im Quellsatz ein Wort nach dem anderen bearbeitet, ist es relativ langsam und weist eine begrenzte Parallelisierung auf (Koehn 2017). Um diese Schwächen zu umgehen, wendeten Vaswani u. a. (2017) einen Self-Attention-Mechanismus (auch als Intra-Attention bezeichnet) an. Self-Attention zeigte Erfolg in anderen Bereichen, wie zum Beispiel in dem Textuellen Schließen (textual entailment, TE) und dem Leseverständnis. Bei dem Self-Attention-Mechanismus werden verschiedene Positionen einer einzelnen Sequenz verknüpft, um eine Repräsentation der Sequenz zu berechnen (ebd.). Bei diesem Mechanismus wird daher die Zuordnung zwischen einem Input-Wort und einem anderen Input-Wort – nicht nur zwischen einem Input-Wort und einem Output-Wort – berechnet, indem mehr Kontext jedes Input-Wortes berücksichtigt wird, um es zu disambiguieren (Koehn 2017).

Mehrere Studien zeigen, dass der NMÜ-Output besser als der PBMÜ-Output ist: Toral & Sanchez-Cartagena (2017) untersuchten neun Sprachkombinationen mit einem NMÜ-System und mehreren PBMÜ-Systemen und fanden heraus, dass das NMÜ-System eindeutig das beste PBMÜ-System für alle Sprachrichtungen vom Englischen in eine andere Sprache übertrifft. Hierbei wurden die Parameter Ähnlichkeit der Outputs, Flüssigkeit (fluency), Wortstellung sowie die Auswirkung von Satzlänge und Performance auf verschiedene Fehlerkategorien zum Vergleich der Systeme herangezogen. Der NMÜ-Output unterschied sich wesentlich und war flüssiger und genauer hinsichtlich der Wortstellung im Vergleich zu dem Output der PBMÜ-Systeme. Zudem waren die NMÜ-Systeme präziser bei der Erzeugung von flektierten Formen. Bentivogli u. a. (2016) beschäftigten

sich mit der Analyse des PE-Aufwands und der Fehlertypen bei der Übersetzung aus dem Englischen ins Deutsche. Im Vergleich zur PBMÜ wies die NMÜ weniger PE-Aufwand, weniger Morphologiefehler und wesentlich weniger Wortstellungsfehler auf. Insbesondere bei einem Sprachenpaar, das sich bezüglich der morphologischen Fülle (morphological richness) und Wortstellung unterscheidet, ist ein solches Ergebnis bedeutsam. Für das in der vorliegenden Studie untersuchte Sprachenpaar (Deutsch-Englisch) fand Popović (2017) (sowie Popović 2018) für beide Übersetzungsrichtungen heraus, dass die NMÜ im Allgemeinen weniger Probleme aufwies. Die genauen Stärken der NMÜ gegenüber der PBMÜ lagen in dem Umgang mit den folgenden Punkten: (1) Verbstellung, Verbform und Vermeiden von Verbauslassungen; (2) englischen nominalen Kollokationen und deutschen Komposita; (3) Artikeln; (4) Phrasenstruktur (ebd.).

In letzter Zeit hat die NMÜ weitere Fortschritte bei der Übersetzung auf Dokumentebene erzielt, in der Kontextinformationen aus mehreren Sätzen im Umfeld des aktuellen Satzes oder aus dem ganzen Dokument genutzt werden, um die Übersetzungsqualität des aktuellen Satzes sowie die Konsistenz, Kohärenz und Kohäsion im übersetzten Dokument zu verbessern (Zhang & Zong 2020). In diesem Bereich konzentrierte sich kürzlich eine Forschungsgruppe (Voita u. a. 2018; 2019) auf die Entwicklung von NMÜ-Ansätzen, die eine bessere Übersetzung auf Dokumentebene bei dem Sprachenpaar Englisch-Russisch ermöglichen. Sie verwendeten nur monolinguale Zieldaten und zeigten eine Verbesserung bei Diskursphänomenen wie Deixis, Ellipsen und lexikalischer Kohäsion (Voita u. a. 2019). Zudem konnten Voita u. a. (2018) gezielt für die Anaphora-Auflösung mithilfe eines kontextfähigen NMÜ-Modells (context-aware model) basierend auf einer Transformer-Architektur Verbesserungen insbesondere im Falle von ambigen Pronomen beobachten. Darüber hinaus zeigten die Ergebnisse, dass das Modell Anaphora-Relationen induziert (ebd.). Mit Fokus auf einer korrekten Übersetzung von Pronomen haben Müller u. a. (2018) mehrere kontextfähige NMÜ-Modelle mit einem großen Datensatz für das Sprachenpaar Englisch-Deutsch getestet und belegten eine Verbesserung bei der Übersetzung von Pronomen mithilfe von Multi-Encoder-Architekturen. Ebenfalls für das Sprachenpaar Englisch-Deutsch untersuchten Stojanovski & Fraser (2018; 2019) Diskursphänomene mithilfe von trainierten kontextfähigen NMÜ-Modellen und konnten genauere Übersetzungen von Pronomen sowie eine höhere Kohärenz im Vergleich zum Baseline-Modell realisieren. Sowohl für Deutsch-Englisch als auch für Englisch-Russisch konnte in einer weiteren Studie (Matusov 2019) durch die Berücksichtigung des Kontexts der vorherigen Sätze oder des ganzen Dokuments eine konsistente NMÜ sowie eine bessere Pronomenauflösung erreicht werden, wobei die Qualitäts-

verbesserung beim Sprachenpaar Deutsch-Englisch höher war als bei Englisch-Russisch.

Durch die Berücksichtigung des Kontexts in der NMÜ erarbeiten weitere Studien verschiedene Strategien für die Übersetzung von Mehrwortausdrücken (Multiword Expressions, MWEs). Ein MWE ist ein Ausdruck, der aus zwei oder mehr Wörtern besteht, die sich als Einheit verhalten; Beispiele hierfür sind Idiome, Funktionsverbgefüge, Verb-Partikel-Konstruktionen, Komposita und Mehrwortentitäten (Constant u. a. 2017). Eine Verbesserung bei der Übersetzung von MWEs zeigen Zaninello & Birch (2020) bei dem Sprachenpaar Englisch-Italienisch, die durch Annotation sowie Datenerweiterung (data augmentation) mithilfe externer sprachlicher Ressourcen erreicht wurde. In einer weiteren Studie führten Gamallo & Garcia (2019) unter Verwendung eines unbeaufsichtigten NMÜ-Ansatzes (unsupervised NMT) die Übersetzung als einen Prozess der Wortkontextualisierung durch, indem lexikosyntaktische Kontexte und Auswahlpräferenzen berücksichtigt werden und zeigten eine Verbesserung bei der Übersetzung von MWEs aus dem Englischen ins Spanische. Auch Rikters & Bojar (2019) testeten weitere Strategien zur Übersetzung von MWEs und konnten die Übersetzung von englischen MWEs ins Lettische und Tschechische verbessern, indem sie zweisprachige Paare von MWE-Kandidaten dem parallelen Korpus hinzufügten, das zum Trainieren des NMT-Systems verwendet wurde. Diese Studien liefern einen kurzen Überblick über einige NMÜ-Fortschritte der letzten Zeit, bevor die Schwächen dieses Ansatzes genauer betrachtet werden.

In Bezug auf die Schwächen der NMÜ im Vergleich zu der PBMÜ nennt Popović (2017) (sowie Popović 2018) für das untersuchte Sprachenpaar (Deutsch-Englisch) Folgendes: Die dominanten Probleme der NMÜ befanden sich in der Übersetzung von Präpositionen, englischen ambigen Wörtern ins Deutsche und bei der Bildung der Verlaufsform bei englischen Verben. Gleichzeitig stellte die Übersetzung von Präpositionen ein Hindernis sowohl für den NMÜ- als auch den PBMÜ-Ansatz dar (ebd.). Ferner wird für unterschiedliche Sprachenpaare die schlechte Übersetzung von langen Sätzen in mehreren Studien thematisiert (Bentivogli u. a. 2016; Koehn 2017; Toral & Sanchez-Cartagena 2017). Dieses Problem wurde mit der Einführung der Attention-Modelle einigermaßen behoben. Bis zu einer Satzlänge von 60 Wörtern ist der NMÜ-Output besser als der der SMÜ. Bei längeren Sätzen übertreffen die SMÜ-Systeme. (Koehn 2017) Bentivogli u. a. (2016) berichten von einer weiteren Systemschwäche bei der Neuordnung bestimmter linguistischer Konstituenten, die ein tiefes semantisches Verständnis erfordern. Weitere Schwächen der NMÜ fasste Koehn (2017) zusammen, in der er die NMÜ- und PBMÜ-Ansätze unter die Lupe nahm:

- Erstens haben NMÜ-Systeme, die auf Subwort-Ebene funktionieren, Schwierigkeiten bei der Übersetzung seltener Wörter, obwohl der NMÜ-Output bei seltenen Wörtern besser als der der PBMÜ ist. Dies kann insbesondere bei stark flektierten Sprachen ein Problem darstellen, da viele Flexionsformen selten auftreten können.
- Zweitens haben NMÜ-Systeme eine steilere Lernkurve in Bezug auf die Trainingsdatenmenge. Dies führt zu einer schlechteren Qualität (im Vergleich zu PBMÜ-Systemen) bei Low-Ressource-Settings, jedoch zu einer höheren Qualität bei High-Ressource-Settings.
- Drittens erfüllt das Attention-Modell nicht immer die Rolle eines Wort-Alignment-Modells. In SMÜ-Systemen bieten die Alignments nützliche Debug-Informationen, um das Modell zu inspizieren. Das Attention-Modell hat eine breitere Rolle, z. B. werden beim Übersetzen eines Verbs das Subjekt und Objekt zur Disambiguierung berücksichtigt.
- Viertens – die Aufgabe der Dekodierung besteht darin, die vollständige Satzübersetzung mit der höchsten Wahrscheinlichkeit zu finden. In NMÜ-Systemen ermöglicht die Beam-Suche-Dekodierung eine verbesserte Übersetzungsqualität nur bei begrenzten Beams. Bei einem größeren Suchraum verschlechtert sich die Qualität.
- Fünftens ist die Qualität des NMÜ-Outputs gering bei der Übersetzung domänenspezifischer Texte, da NMÜ-Systeme die Adäquatheit (adequacy) für die Flüssigkeit (fluency) opfern. Dementsprechend kann der Output z. B. bei einem generischen NMÜ-System in Fachbereichen wie Recht zwar flüssig, aber inadäquat, ausfallen. In einem solchen Fall liefern PBMÜ-Systeme eine bessere Übersetzung. (Koehn 2017) Da Letzteres für die vorliegende Studie von besonderer Bedeutung ist, soll darauf im Folgenden näher eingegangen werden.

In der vorliegenden Studie wurden generische Systeme für die Übersetzung in der technischen Domäne verwendet. Während eine Domänenadaptation bzw. Terminologieintegration in der RMBÜ, SMÜ und HMÜ mithilfe fachspezifischer Wörterbücher bzw. durch das Training mittels fachspezifischer Parallelkorpora möglich gewesen wäre, befand sich die Domänenadaptation bzw. Terminologieintegration in der NMÜ zur Zeit der Durchführung der Studie noch in der experimentellen Phase (vgl. Eisold 2017). Damit die Studie auf einer einheitlichen Basis durchgeführt wird, wurden alle Systeme in ihrem Ist-Zustand, d. h. ohne Terminologieintegration oder Training mit domänenspezifischen Daten, verwendet.

Die Problematik der Terminologieübersetzung wurde in der Studie umgangen, indem die spezifischen Termini in den analysierten Sätzen durch geläufige Begriffe ersetzt wurden (Genaueres dazu unter §4.5.3.1, Schritt [4]).

Die Verwendung von Terminologien im Trainingskorpus eines NMÜ-Systems ist kritisch, da jede Erweiterung des Vokabulars die notwendige Rechenleistung und Speichernutzung deutlich erhöht. Ein Wort, das im Vokabular des NMÜ-Systems nicht enthalten ist, wird als OOV-Wort (out of vocabulary) betrachtet und in der Regel unübersetzt oder als Label ausgegeben. Problematisch ist, dass das System anhand der Frequenz über das NMÜ-Vokabular entscheidet. Da Fachtermini nicht unbedingt häufig vorkommen, werden sie unter Umständen wie OOV-Wörter behandelt. (Eisold 2017: 118f.) In seinen ersten Experimenten passte Systran sein generisches NMÜ-System an eine spezifische Domäne an, indem es zusätzliche Trainingsepochen mittels neu verfügbarer In-Domain-Daten – nach Abschluss des grundlegenden Systemtrainingsprozesses – durchführte (Crego u. a. 2016). Diese inkrementelle Anpassung wurde in wenigen Sekunden ausgeführt und zeigte Qualitätsverbesserungen bei den In-Domain-Sets. Crego et al. stellten gleichzeitig fest, dass das vollständige Training (anstatt der inkrementellen Anpassung) zu einer zusätzlichen Verbesserung der Genauigkeit führte, da das In-Domain-Vokabular in das neue vollständige Modell aufgenommen wurde. Allerdings nahm das vollständige Training 17 Stunden in Anspruch. (ebd.) Außerdem wurde diese Methode aufgrund ihres Flexibilitätsmangels kritisiert (Dinu u. a. 2019). Dadurch, dass die Entitäten in dieser Methode durch spezielle Tags ersetzt werden, ersetzt das Modell den Platzhalter unabhängig vom grammatischen Kontext immer durch dieselbe Phrase (ebd.). Anders als bei einem PBMÜ-Decoder, dem die Reihenfolge der zu übersetzenden Quellphrasen bekannt ist, verfügt der NMÜ-Decoder nicht über diese Informationen (Chatterjee u. a. 2017). Dies erschwert die Integration von Teilübersetzungen aus externen Ressourcen (wie z. B. zweisprachigen Wörterbüchern) in den NMÜ-Workflow. Eine weitere Schwierigkeit bei der NMÜ-Architektur – im Gegensatz zur Decodierung in der PBMÜ – besteht darin, dass die NMÜ-Architektur keine „coverage constraint" auf die Quellwortpositionen anwendet (d. h. es gibt keine Garantie, dass das System jedes Quellwort genau einmal übersetzt). (ebd.)

Trotz dieser Schwierigkeiten wurden zuletzt einige Ansätze zur Domänenadaptation bzw. Terminologieintegration in der NMÜ entwickelt, die einen Fortschritt bei der konsistenten Übersetzung spezifischer Terminologien nach festgelegten externen Terminologielisten (bzw. Terminologiedatenbanken) belegen (Chatterjee u. a. 2017; Hasler u. a. 2018; Dinu u. a. 2019). Diese Entwicklung wird im Folgenden näher betrachtet.

Chatterjee u. a. (2017) entwickelten eine Methode zur Erweiterung eines vorhandenen NMÜ-Decoders, indem der Übersetzungsprozess anhand von externen übersetzten Terminologielisten (Einschränkungen in Form von XML-Annotationen der Quellwörter zusammen mit den entsprechenden Übersetzungen) gesteuert wird. Durch diese Steuerung werden eine konsistente Übersetzung und korrekte Wortstellung der Terminologie, einschließlich der Fälle externer Wörter, die dem Modell unbekannt sind (OOV), sichergestellt. Somit konnte diese Methode die MÜ-Qualität (i. S. v. höheren AEM-Scores) gegenüber der des Baseline-NMÜ-Decoders signifikant verbessern. (ebd.) Hasler u. a. (2018) führten einen Ansatz zur NMÜ-Decodierung mit terminologischen Einschränkungen unter Verwendung von Decoder-Attentions ein, der sowohl zieltextseitige Einschränkungen als auch Einschränkungen des entsprechenden Quelltexts unterstützt. Dieser Ansatz ermöglichte eine deutliche Reduzierung der Duplizierung (sog. „output duplication“)[7] und der Fehlplatzierung (sog. „constraint placement“)[8] der übersetzten Termini. Die Ergebnisse für vier Sprachenpaare zeigen, dass kundenspezifische Terminologien während der NMÜ-Decodierung eingehalten werden konnten, während die Gesamtübersetzungsqualität (i. S. v. signifikant höheren BLEU-Scores) stieg. Gleichzeitig konnte dieser Ansatz die Rechenkomplexität reduzieren und somit zur schnelleren Decodierung führen. (ebd.) In einem weiteren Ansatz trainierten Dinu u. a. (2019) eine generische NMÜ-Architektur direkt zu lernen, wie sie externe Terminologieeinträge verwendet, die zur Laufzeit bereitgestellt werden. Im Vergleich zur eingeschränkten Dekodierung („constrained decoding“) konnten nach diesem Ansatz in einigen Fällen morphologische Varianten von Terminologieübersetzungen, die von der Terminologiedatenbank bereitgestellt werden, generiert werden. Ferner verzeichnete dieser Ansatz eine höhere Verwendungsrate der Terminologie, kürzere Dekodierungszeit sowie bessere BLEU-Scores als die der Baseline und der eingeschränkten Dekodierung. (ebd.)

Die Ergebnisse der dargestellten Studien belegen den signifikanten Fortschritt der Domänenadaptation bzw. Terminologieintegration in der NMÜ und die damit verbundene Steigerung der Gesamtübersetzungsqualität. Ferner hat das Unternehmen Intento Ende 2019 eine Reihe von MÜ-Systemanbietern aufgelistet,

[7] „Output duplication“ kommt vor, wenn der zu übersetzende Terminus sowohl vom System als auch auf Basis der Terminologie-Einschränkungen (z. B. einer Terminologieliste) übersetzt wird. Diese Duplizierung konnte der Ansatz von Hasler u. a. (2018) mithilfe der Attentions verhindern.

[8] Das Ziel bei der „constraint placement“ ist eine korrekte Platzierung der Übersetzung des Terminus, die Hasler u. a. (2018) mithilfe der quelltextseitigen Einschränkungen zusammen mit den Attentions realisieren konnten.

deren NMÜ-Systeme eine automatische bzw. manuelle Domänenadaption unterstützen (Bruckner 2020: 44).[9] Neben den bekannten MÜ-Anbietern (wie Google, Systran und SDL) existieren mehrere mittelständische Anbieter, deren Computerlinguisten Open-Source-Frameworks nutzen, um NMÜ-Engines kunden- bzw. domänenspezifisch aufzubauen und zu optimieren. Beispiele für diese Anbieter sind Tilde (aus Lettland), Iconic (aus Irland), TextShuttle (aus der Schweiz), Globalese (aus Ungern) und Omniscien (aus Großbritannien). (ebd.: 46) Auf dem deutschen Markt hat DeepL im Mai 2020 von der neusten Systementwicklung der Integration von individuellen Glossaren in sein NMÜ-System berichtet. Diesem Bericht zufolge unterstützt DeepL aktuell die Integration von Glossaren für vier Sprachenpaare und passt die Übersetzung der Glossarbeiträge sowohl grammatisch als auch in der Formulierung an. (DeepL 2020)

Das NMÜ-System, das in dieser Studie untersucht wurde, ist *Google Translate*.[10] Google veröffentlichte 2016 unter dem Titel „Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation" einen Aufsatz, in dem es über seinen Wechsel von dem SMÜ-Ansatz zu dem NMÜ-Ansatz schrieb und das Erreichen eines Meilensteins feierte, dass „[i]n some cases human and GNMT translations are nearly indistinguishable [...] our system's translation quality approaches or surpasses all currently published results" (Wu u. a. 2016: 20). Das Modell von Google Translate besteht aus den drei Komponenten: Encoder-Netzwerk, Decoder-Netzwerk und Attention-Netzwerk. Der Encoder transformiert den Ausgangssatz in eine Liste von Vektoren, ein Vektor pro Input-Symbol. Auf Basis dieser Vektorenliste produziert der Decoder ein Symbol nach dem anderen, bis das spezielle Satzendsymbol (end-of-sentence symbol, EOS) produziert ist. Das Attention-Modul verbindet den Encoder mit dem Decoder und ermöglicht dem Decoder, sich während des Decodings auf verschiedene Stellen des Ausgangssatzes zu fokussieren. (ebd.)

Als Teil des Deep-Learning-Konzepts können NMÜ-Systeme durch die Übersetzung von einer Sprache in eine andere weiterlernen und sich verbessern. Zudem fanden die Google-Translate-Forscher heraus, dass das System in der Lage ist, zwischen zwei Sprachen zu übersetzen, die es vorher nie lernte. In dem Konzept, das Google als „Zero-shot translation" bezeichnet, zeigt Google (Johnson u. a. 2016), dass das System zwischen Japanisch <> Koreanisch übersetzen konnte,

[9] Intento ist ein Unternehmen, „das MÜ-Systeme von Drittanbietern über eine zentrale Schnittstelle in CAT- und TMS-Lösungen so einbindet, dass textsortenbezogen jeweils das geeignetste System genutzt wird" (Bruckner 2020: 44). Die Systemanbieter sind in der Abbildung „Machine Translation Landscape" unter https://blog.inten.to/november-2019-mt-landscape-enterprise-mt-hub-and-lots-of-mt-news-ed25cf6235b7onlineverfügbar [abgerufen am 20.04.2020].

[10] Links zum System Google Translate: https://translate.google.de/

nachdem es zuvor nur zwischen Japanisch <> Englisch und Koreanisch <> Englisch übersetzt hatte. Hierbei schafft das System zunächst eine künstliche Sprache, übersetzt aus der Ausgangssprache in die künstliche Sprache und schließlich aus der künstlichen Sprache in die Zielsprache. Gleichzeitig gab Google 2016 zu, dass die Probleme der MÜ mit der NMÜ keineswegs gelöst seien, so arbeite es an Problemen wie dem Weglassen von Wörtern und der falschen Übersetzung von seltenen Begriffen oder Eigennamen sowie an der Berücksichtigung von erweiterten Kontexten auf Absatz- oder sogar Seitenebene weiter. (vgl. Le & Schuster 2016)

Im folgenden Abschnitt werden MÜ-Evaluationsstudien im Bereich der KS diskutiert. Diese Studien decken verschiedene Evaluationsmethoden sowie die RBMÜ-, SMÜ- und HMÜ-Systeme ab. Nach bestem Wissen der Forscherin wurden bisher keine Evaluationsstudien für NMÜ-Systeme in Zusammenhang mit der KS durchgeführt.

3.4 MÜ-Qualitätsevaluation

Eine zusammenfassende Aussage von Yorick Wilks reflektiert den Forschungsstand der MÜ-Evaluation mit den Worten, dass „more has been written about machine translation evaluation than about machine translation itself“ (King u. a. 2003: 1). Parallel zur Entwicklung der MÜ-Systeme und ihrer Ansätze laufen sowohl in der Forschung als auch in der Industrie fortwährend Evaluationen mit verschiedenen Zielsetzungen auf unterschiedlichen Ebenen, für mehrere Zielgruppen und dementsprechend nach diversen Vorgehensweisen. Zweifellos sind diese Evaluationen von großer Bedeutung für die Entwicklung der MÜ: Mithilfe der Evaluationen können die zugrundeliegenden Probleme oder Fehler erfasst und behoben werden, wodurch einzelne Systeme weiterentwickelt, mehrere Systeme verglichen und MÜ-Ansätze verfeinert oder kombiniert werden können, mit dem Endziel, den MÜ-Output zu optimieren. Eine MÜ-Evaluation erfolgt in der Regel über die Qualitätsbewertung des Outputs eines MÜ-Systems. In diesem Abschnitt werden die Problematik der Qualitätsdefinierung im Rahmen der MÜ-Evaluation thematisiert, die Konstruktion eines Evaluationsdesigns diskutiert und die verschiedenen Evaluationsmethoden zusammen mit einer Literaturübersicht dargestellt.

3.4.1 Qualität der MÜ

Eine Qualitätsmessung ist in jeder Disziplin ein komplexes Thema. In diesem Sinne stellt eine Auswertung der MÜ-Qualität keine Ausnahme dar. Diese Pro-

blematik wurde in mehreren Arbeiten wie folgt verdeutlicht:

> In reality quality is whatever the customer wants it to be. This in itself demonstrates just how diverse and heterogeneous quality standards and all aspects of translation quality must be and have always been. (Burchardt & Harris 2017: 128)
>
> Theorists and professionals overwhelmingly agree there is no single objective way to measure quality. Yet every day, translators, editors, revisers, clients and many others nonetheless have to do just this. (Drugan 2013: 35)
>
> A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs. (Koby u. a. 2014: 416)

Diese Zitate sowie die zahlreichen existierenden wissenschaftlichen Definitionen der MÜ-Qualität zeigen, dass die MÜ-Qualitätsevaluationen variieren und primär von ihrem Zweck sowie ihrer Interessentengruppe abhängen (vgl. Hutchins 1997; White 2003). Diese Faktoren bestimmen die Evaluationsvorgehensweise, inkl. der Evaluationsmethode, ihrer Ebene, ihrer Komplexität sowie ihres Umfangs.

Folgende Beispiele verdeutlichen den Einfluss der genannten Faktoren auf die MÜ-Qualitätsevaluation: Ein Tourist, der mithilfe seiner Übersetzungsapp versucht, während seiner Reise zur simplen Verständigung ein paar Sätze zu entziffern, kann mit der Übersetzungsqualität seiner App zufrieden sein, wenn er in dieser Situation schnell Hilfe und Problemlösungen findet, selbst wenn die erhaltene Satzstruktur nicht perfekt sein sollte. Man spricht hier von einer Informativübersetzung (gisting translation), die ein rudimentäres Verständnis vom fremdsprachlichen Text zum Ziel hat. Anders als ein Projektleiter auf einer Dienstreise, der versucht, einen kurzen Bericht von seinem Übersetzungsprogramm übersetzen zu lassen. Seine Erwartungen bezüglich der Korrektheit der Übersetzung können hier entsprechend höher liegen. Ein letztes Bespiel ist von einem Unternehmen, das sämtliche standardisierten Dokumentationen oft in mehrere Fremdsprachen übersetzen muss. Bei diesem Unternehmen würde die Kostenreduzierung eine große Rolle spielen (z. B. mit welchem MÜ-System die Übersetzungbzw. Post-Editing-Kosten sich reduzieren lassen). Anhand dieser Beispiele ist zu erkennen, dass die MÜ-Qualitätsdefinition sowie das Evaluationsdesign unterschiedlich sein können und von verschiedenen Faktoren abhängig sind. Im Folgenden beschäftigen wir uns mit der MÜ-Qualitätsdefinition in der Forschung

und im Abschnitt §3.4.2 wird detailliert auf das MÜ-Evaluationsdesign eingegangen.

In der MÜ-Evaluationsliteratur sind zahlreiche Definitionen der MÜ-Qualität zu finden. Diese Definitionen decken grundsätzlich drei Qualitätskriterien (Genauigkeit, Verständlichkeit und Stil) ab (vgl. Hutchins & Somers 1992: 164), die in den Studien unterschiedlich bezeichnet und gruppiert werden. Somit wird man in der MÜ-Evaluationsforschung mit zahlreichen Synonymen von Qualitätskriterien bzw. sich überlappenden Kriterien konfrontiert. Die „intelligibility" ist ein Bestandteil vieler Studien, die entweder zusammen mit der „fidelity", wie es in der Methode von DARPA[11] der Fall ist (vgl. White 2003) oder zusammen mit der „accuracy" (vgl. Arnold 1994: 169) zur Bewertung der MÜ-Outputsqualität herangezogen wird. Andere Studien bevorzugen die Verwendung der Bezeichnung „fluency" (anstelle von „intelligibility") und der Bezeichnung „adequacy" (anstelle von „fidelity") (vgl. Hamon 2007). Das Framework for the Evaluation of Machine Translation in ISLE (FEMTI) bewertet die Qualität anhand der Kriterien „comprehensibility", „readability", „style" und „clarity" (vgl. King u. a. 2003). Auch in FEMTI kommt die „intelligibility" vor, allerdings unter der Verwendung eines Synonyms, nämlich „comprehensibility". Dennoch unterscheidet T. C. Halliday (Van Slype 1979: 62) in seiner Analyse zwischen der „intelligibility" und der „comprehensibility". Van Slype (ebd.) kombiniert hingegen die „comprehensibility" und „clarity" in einer Definition für die „intelligibility". In einer weiteren Studie führen Vanni & Miller (2002) die „comprehensibility", „readability", „style" und „clarity" zu einer Qualitätsmetrik bezeichnet als „clarity" zusammen. Im selben Jahr veröffentlichte das Linguistic Data Consortium (LDC 2002) eine häufig zitierte Evaluationsstudie, in der die Qualität anhand der „adequacy" und „fluency" gemessen wird. Um eine Gruppierung der Kriterien zu vermeiden, definierte Coughlin (2003) eine Skala für die „acceptability" ohne Spezifizierung von Qualitätskriterien. In der Skala verbergen sich jedoch die Kriterien Richtigkeit der Grammatik, Genauigkeit und Verständlichkeit. Vor diesem Labyrinth fassten Hutchins & Somers (1992: 164) die Qualitätskriterien in Genauigkeit, Klarheit und Stil zusammen und berücksichtigten in ihren Definitionen die möglichen Synonyme wie folgt:

> (a) Fidelity or accuracy, the extent to which the translated text contains the ‚same' information as the original; (b) Intelligibility or clarity, the ease with which a reader can understand the translation; and (c) Style, the extent to which the translation uses the language appropriate to its content and intention. (Hutchins & Somers 1992: 164)

[11] US Defense Advanced Research Projects Agency (DARPA)

Hutchins und Somers lieferten damit umfassende und klar aufgeteilte Qualitätskriterien (vgl. Fiederer & O'Brien 2009). Daher werden sie bei der Humanevaluation in der vorliegenden Studie herangezogen (siehe §4.5.5.1).

Die Idee, dass Qualität mit der Kundenzufriedenheit erreicht wird, und die Erkenntnis, dass die Anforderungen an die Übersetzungsqualität je nach Inhalt, Zweck und Zielgruppe variieren, führten zur Entwicklung von zwei dynamischen bzw. flexiblen Frameworks zur Qualitätsevaluation: das Dynamic Quality Framework (DQF) und das Multidimensional Quality Metrics (MQM). Im Jahr 2011 hat TAUS das DQF mit dem Ziel entwickelt, die Evaluation der Übersetzungsqualität zu standardisieren (Attila 2014). Mit dem gedanklichen Ansatz, dass es keine „one-size-fits-all"-Methode bzw. Vorgehensweise zur Evaluation der Übersetzungsqualität gibt, stellt das DQF eine Plattform mit „rich knowledge base on Quality Evaluation with best practices, reports, templates and a number of tools" zur Evaluation von Human- und maschineller Übersetzung zur Verfügung (ebd.). Die Evaluation umfasst mehrere Vorgehensweisen: Vergleich von Übersetzungen, Bewertung ihrer „accuracy" und „fluency", Messung der Post-Editing-Produktivität sowie Scoring der Übersetzungssegmente auf Basis einer Fehlertypologie (ebd.). Damit bietet das DQF dem Bewerter alle erforderlichen Mittel, um das/die für seine spezifischen Qualitätsanforderungen am besten geeignete(n) Evaluationsmodell(e) bzw. -metriken auszuwählen. Das zweite flexible Framework ist MQM, entwickelt vom Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) im Jahr (Lommel u. a. 2013). MQM ermöglicht eine Evaluation anhand eines umfangreichen Katalogs von Fehlertypen, die in einer Hierarchie angeordnet sind (ebd.); näheres dazu unter §3.4.3.1 „Methoden der Humanevaluation". Im Jahr 2014 begannen TAUS und DFKI eine Harmonisierung von DQF und MQM mit dem Ziel, die Lücke zwischen den Definitionen und Spezifikationen der beiden Modelle zu schließen (Attila 2014).

3.4.2 Evaluationsdesign

Wie im vorherigen Abschnitt diskutiert wurde, wird die MÜ-Evaluation von verschiedenen Faktoren beeinflusst. Dies wiederum lässt ein Evaluationsdesign wie ein Puzzlebild betrachten, dessen Teile sich aus Faktoren wie der Interessentengruppe, dem Evaluationsfokus, dem/den analysierten MÜ-System(en) und seinem/ihrem Analysemodus, der durchgeführten Datenanalyse, dem Evaluationsmaterial sowie den Evaluationsteilnehmern zusammensetzen (vgl. Hutchins 1997; Weber 1998; White 2003). Im Folgenden werden diese Faktoren zusammen mit den damit verbundenen möglichen Evaluationsdesigns dargestellt:

3.4.2.1 Interessentengruppe der Evaluation

Mehrere Studien beschäftigten sich mit der MÜ-Evaluation aus Perspektive der Interessentengruppe (vgl. Hutchins 1997; Weber 1998; White 2003), da diese den Fokus und das Design der Evaluation weitgehend bestimmen. Folgende Interessentengruppen werden häufig genannt: (monolinguale) Nutzer, Übersetzer bzw. Übersetzungseditoren, Manager bzw. Käufer eines Systems, Entwickler bzw. Systemanbieter sowie Forscher (ebd.). Weber (1998: 65) unterscheidet zwischen direkten und indirekten Nutzern. Er betrachtet die Unternehmer und Übersetzer als direkte Nutzer und klassifiziert sie zusammen mit den Forschern (ebd.). Auf der anderen Seite bezeichnet er die Übersetzungsrezipienten als indirekte Nutzer und klassifiziert sie zusammen mit den Entwicklern (ebd.). Eine detailliertere, aber ähnliche Klassifizierung bietet White (2003: 209ff.), indem er die Ziele der folgenden Zielgruppen ausführlich vergleichend darstellt: (1) Endnutzer mit einer Unterscheidung zwischen Übersetzern, Übersetzungseditoren und monolingualen Nutzern; (2) Manager mit einer Unterscheidung zwischen Betriebsleitern und Beschaffungsmanagern; (3) Entwickler mit einer Unterteilung in Forscher und Produzenten; (4) Anbieter; und (5) Investoren mit einer Unterteilung in Forschungsorganisationen und Spezialfinanzierer. Hutchins (1997: 418) differenziert in seiner Aufteilung der Interessentengruppen zwischen potenziellen Nutzern, Systemkäufern, Forschern und Entwicklern; und verknüpft diese mit dem Fokus der verschiedenen Evaluationen, wie im folgenden Punkt erläutert wird.

In dieser Studie wird die Evaluation von Übersetzern für Forschungszwecke durchgeführt.

3.4.2.2 Fokus der Evaluation

Jede der obengenannten Zielgruppen hat bei der Evaluation einen bestimmten Fokus. Im Folgenden werden drei Aufteilungen der Evaluationen nach Fokus dargestellt:

3.4.2.2.1 Adäquanz- vs. Diagnostik- vs. Fortschrittsevaluation

Eine der Evaluationsaufteilungen nach Fokus ist die Aufteilung in die drei Typen: Adäquanz- vs. Diagnostik- vs. Fortschrittsevaluation (auch Performanzevaluation) (Hutchins 1997: 418; Weber 1998: 65): In einer *Adäquanzevaluation* wird untersucht, ob der MÜ-Output seinen Zweck aus Sicht der Nutzer bzw. Käufer erfüllt, wodurch die Systemeignung in einem bestimmten Betriebskontext beurteilt werden kann. Bei einer *Diagnostikevaluation* zielen die Entwickler und Forscher darauf ab, die Systemfehler und -limitationen zu identifizieren, die Fehlerursachen

aufzudecken, die Fehler zu beheben und entsprechend das System zu optimieren. Durch eine *Fortschrittsevaluation* (Performanzevaluation) messen die Entwickler bzw. die Forscher den Fortschritt eines Systems nachdem bestimmte Anpassungen vorgenommen bzw. neue technische Implementierungen durchgeführt wurden und vergleichen somit seine Leistung in verschiedenen Entwicklungsphasen. (Hutchins 1997: 418)

Entsprechend dieser Unterteilung handelt es sich bei der vorliegenden Studie um eine Adäquanzevaluation.

3.4.2.2.2 Produktorientierte vs. prozessorientierte Evaluation

Eine *produktorientierte Evaluation* hat die Bewertung des MÜ-Outputs im Fokus. Die Evaluation wird anhand von tatsächlichen Daten und somit retrospektiv zu diagnostischen Zwecken durchgeführt. Eine *prozessorientierte Evaluation* hingegen befasst sich mit der Systemfunktionalität. Hierbei steht die Funktionalität bei zukünftigen Übersetzungen im Fokus und damit erfolgt die Analyse prospektiv für prognostische Zwecke. (Weber 1998: 64f.)

Nach dieser Unterteilung geht es bei dieser Studie um eine produktorientierte Evaluation.

3.4.2.2.3 Deklarative, typologische vs. operationale Evaluation

Nach Weber (ebd.) befasst sich eine *deklarative Evaluation* mit der Qualität des MÜ-Outputs. Diese erfolgt in der Regel mithilfe von Humanbewertern, die verschiedene Aspekte der Qualität, z. B. Genauigkeit, Verständlichkeit usw. beurteilen. Die deklarative Evaluation ist insbesondere für Forscher und direkte Nutzer von Interesse. Eine *typologische Evaluation* hingegen beschäftigt sich mit den linguistischen Phänomenen, die das System abdeckt, sprich mit der „linguistic coverage“ bzw. Kompetenz des Systems. Diese Art von Evaluationen wird in der Regel von den Systementwicklern durchgeführt. In einer *operationalen Evaluation* steht die Kosteneffektivität im Mittelpunkt und wird entweder in situ oder durch den Einsatz eines Testkorpus bewertet. (ebd.)

Gemäß dieser Aufteilung handelt es sich bei der vorliegenden Studie um eine deklarative Evaluation, die durch den Einsatz von drei Methoden die Qualität des MÜ-Outputs bewertet.

3.4.2.3 MÜ-System(e): wie, was und wie viele evaluiert werden

3.4.2.3.1 Black-Box- vs. Glas-Box-Evaluation

Je nachdem, ob bei der Evaluation die systeminternen Abläufe analysiert werden, unterscheidet man zwischen (Weber 1998: 64; White 2003: 215) einer Black-Box-Analyse und einer Glass-Box-Analyse. In der *Black-Box-Analyse* stehen dem Bewerter keine Informationen zu den genauen systeminternen Vorgängen zur Verfügung. Der Fokus liegt hierbei nur auf dem Systemoutput und somit kann die exakte Ursache eines Systemfehlers nicht festgestellt werden. Für die Nutzer zum Beispiel sind die Fehlerursachen irrelevant, sie interessieren sich in der Regel nur für den Output. In der *Glas-Box-Analyse* hingegen ist der genaue Systemaufbau bekannt. Die Ergebnisse können in Hinsicht auf die Systemkomponente und -vorgänge analysiert werden. Dementsprechend lassen sich die MÜ-Fehlerursachen interpretieren. Solche Analysen werden vor allem von Systementwicklern durchgeführt.

In dieser Studie handelt es sich um eine Black-Box-Analyse, da der Fokus auf dem Vergleich von den Szenarien des KS-Einsatzes in verschiedenen MÜ-Systemen liegt und somit steht der reine Output im Mittelpunkt.

3.4.2.3.2 Evaluation von einzelnen Komponenten vs. dem ganzen System

In einer Evaluation kann eine bestimmte Komponente im System oder das komplette System untersucht werden (Weber 1998: 65): Untersuchungen einer *bestimmten Systemkomponente* werden in der Regel von Entwicklern durchgeführt, z. B. im Rahmen von Fortschrittsevaluationen zur Optimierung der Systemleistung. Hierbei ist eine Glas-Box-Analyse erforderlich. Auf der anderen Seite interessieren sich die Nutzer hauptsächlich für den Output des *kompletten Systems*, daher geht es bei solchen Evaluationen meist um Black-Box-Analysen.

In dieser Studie werden mehrere Systeme vollständig verglichen. Die Untersuchung von einzelnen Komponenten ist angesichts der Zielsetzung der Studie irrelevant.

3.4.2.3.3 Mikro- vs. Makroevaluation

Mehrere Forscher unterscheiden zwischen einer Mikro- und Makroevaluation (Van Slype 1979; Sager 1994; Weber 1998). Hierbei haben Weber (1998: 64) und Sager (1994: 264f.) eine ähnliche Perspektive: Bei der *Mikroevaluation* wird ein einzelnes System im Detail analysiert, während bei der *Makroevaluation* mehrere MÜ-Systeme untersucht und einander gegenübergestellt werden. Nach Sager (1994: 265) kann eine *Mikroevaluation* mehrere Ziele haben: diagnostische

Zwecke zur Untersuchung von MÜ-Fehlerursachen oder prognostische Zwecke zur Bewertung der Fehler, Arbeit an möglichen Lösungen und zukünftiger Vermeidung. Diese Art von Evaluation wird in der Regel bei der Systementwicklung durchgeführt. Entsprechend sind sie für Entwickler von Interesse und erfordern eine Glass-Box-Analyse. Im Rahmen einer *Makroevaluation* werden konkurrierende Systeme in Bezug auf ihre Fähigkeit zur Erfüllung des von ihnen beabsichtigten Zwecks (z. B. Post-Editing, direkte Verwendung usw.) beurteilt. Eine Makroevaluation wird als erster Schritt zum Vergleich von Zeit- und Kostenfaktoren zwischen der Human- und maschinellen Übersetzung eingesetzt. (ebd.: 264f.) Für Van Slype (1979) steht nicht die Anzahl der untersuchten Systeme im Mittelpunkt. Er berücksichtigt eine Makro- und Mikroevaluation aus der Perspektive ihrer Zielsetzung: Eine *Makroevaluation* ist eine Gesamtevaluation, bei der die Akzeptanz eines Systems gemessen werden kann, zwei Systeme oder zwei Versionen eines Systems verglichen werden können oder die Usability eines Systems bewertet werden kann. Eine *Mikroevaluation* ist eine detaillierte Evaluation zur Bewertung der Verbesserungsfähigkeit eines Systems oder Erstellung einer Verbesserungsstrategie. (ebd.: 12)

In der vorliegenden Studie werden fünf unterschiedliche MÜ-Systeme verglichen; somit handelt es sich nach Weber (1998) um eine Makroevaluation.

3.4.2.4 Datenanalyse

3.4.2.4.1 Quantitative vs. qualitative Evaluation

Weber (1998: 64) beschreibt eine *qualitative* Evaluation als eine tiefgründige Analyse (deep) und im Kontrast dazu die *quantitative* als oberflächliche (shallow) Analyse. Typischerweise werden bei dem *quantitativen* Verfahren zählbare Variablen gemessen bzw. berechnet und statistisch analysiert. Ein besseres Verständnis für die ermittelten quantitativen Ergebnisse ermöglicht das qualitative Verfahren. Bei dem *qualitativen* Verfahren werden die untersuchten Phänomene, z. B. mithilfe von Beobachtungen oder Interviews, tiefgründig analysiert. (vgl.: Creswell & Clark 2007: 415f.; Saldanha & O'Brien 2014: 23)

Da die beiden Verfahren sich ergänzen, werden die Daten in dieser Studie sowohl qualitativ als auch quantitativ nach einem Mixed-Methods-Ansatz analysiert.

3.4.2.4.2 Objektive vs. subjektive Evaluation

Nach Weber (1998: 64) sind automatisierte Evaluationen *objektiv*, während von Menschen durchgeführte Evaluationen als *subjektiv* bezeichnet werden. Hierbei

betrachtet Weber (ebd.) in der vorherigen Aufteilung quantitative Evaluationen als *objektiv*, da sie in der Regel statistisch bzw. mithilfe von Software und programmierten Tools durchgeführt werden. Auf der anderen Seite klassifiziert Weber (ebd.) qualitative Evaluationen als *subjektiv*, da deren Ergebnisse in der Regel durch den Einsatz von Probanden erzielt werden. Doch eine scharfe Trennlinie zwischen automatischen Evaluationen als objektiv und Humanevaluationen als subjektiv lässt sich nicht ziehen, da beide Arten von Evaluationen einen verschiedenen Objektivitätsgrad aufweisen (vgl. Doherty 2017): Bei einer automatischen Evaluation kann auf die Beteiligung von Menschen nicht vollständig verzichtet werden. Zwar werden die Ergebnisse mithilfe von Algorithmen oder mathematisch ermittelt, dennoch werden die Auswahl des Datensatzes, Auswahl des statistischen Tests und Designs des statistischen Tests von Menschen getroffen (z. B. ergeben nicht selten zwei statistische Tests bei der Analyse einer und derselben Fragestellung unterschiedliche quantitative Werte, die den subjektiven Aufbau des jeweiligen Tests reflektieren). Ebenfalls hängt der Objektivitätsgrad einer Humanevaluation von Faktoren wie Anzahl der Probanden und Evaluationssettings ab.

Diese Studie wird mithilfe von mehreren Methoden durchgeführt, wobei die Daten sowohl mithilfe von Humanübersetzern als auch mit automatischen Metriken analysiert werden. Ein adäquater Objektivitätsgrad stützt sich sowohl bei der Evaluation als auch bei der Qualitätssicherung auf der Triangulation der Ergebnisse sowie dem Einsatz von zwei AEMs und mehreren qualifizierten Humanübersetzern.

3.4.2.4.3 Linguistische vs. technische Evaluation

Eine *linguistische* Evaluation unterteilt Weber (1998: 64f.) in zwei Arten: eine positive Messung im Sinne einer Bewertung der Systemleistung (performance) und eine negative Messung der Systemfehler. Für eine *technische* Evaluation liefert Weber keine Erläuterung. Durch die Aufteilung ist davon auszugehen, dass er sich damit auf die technischen Eigenschaften des Systems bezieht, z. B. die Nutzeranzahl, seine Geschwindigkeit, mögliche Erweiterbarkeit usw. Eine gewisse Überschneidung zwischen den beiden Arten liegt auf der Hand.

Diese Studie stellt eine linguistische Evaluation der beiden erwähnten Arten dar. In der durchgeführten Fehleranalyse handelt es sich um eine Fehlermessung. In der daran anschließenden Humanevaluation werden sowohl korrekte als inkorrekte Systemoutputs von Übersetzern bewertet, wodurch die Leistung der Systeme beurteilt werden kann.

3.4.2.5 Evaluationsmaterial

Weber (1998: 64f.) unterscheidet bei dem Evaluationsmaterial zwischen einem *Testkorpus* und einer *Testsuite*.

Bei einem *Korpus* handelt es sich allgemein betrachtet um „a body of naturally occurring language" (McEnery u. a. 2006: 4). Genau genommen stellt ein Korpus authentische Texte dar, die für bestimmte Zwecke zusammengestellt werden und einen bestimmten Texttyp repräsentieren (ebd.: 4f.). Daher spielt die Repräsentativität des Korpus eine wesentliche Rolle bei der MÜ-Evaluation, denn es wird erwartet, dass das Korpus „the full range of variability" in einer Sprachvarietät umfasst (ebd.: 13). Ein wesentlicher Vorteil von Korpora gegenüber Testsuites besteht darin, dass sie „more representative of NLP input in the sense they are real language texts rather than artificially constructed data" sind (Balkan u. a. 1994: 33f.). Dies stellt zugleich eine Schwierigkeit dar, denn natürliche Texte sind aufgrund der Vielfalt der enthaltenen Sprachphänomene strukturell komplex.

Eine *Testsuite* wird als ein „carefully constructed set of inputs, where typically each input is designed to probe the system's behavior with respect to some specific phenomenon" definiert (King 1993). Die Verwendung von Testsuites in der MÜ-Evaluationsforschung begann in den 90er Jahren mit einzelnen bekannten Studien (z. B. King & Falkedal 1990; Isahara 1995; Koh u. a. 2001). Die Entwicklung der NMÜ weckte jedoch in den letzten Jahren erneut das Interesse der Forscher an analytischeren Diagnoseverfahren, mit denen die MÜ-Qualität bei spezifischen Phänomenen untersucht werden kann (Macketanz u. a. 2018). So untersuchen aktuelle Studien auf Basis von spezifischen Testsuites Phänomene wie Pronomen (Guillou & Hardmeier 2016), strukturelle Divergenzen (Isabelle u. a. 2017) und Verbpartikelkonstruktionen (Schottmüller & Nivre 2014). Zudem werden Testsuites in der letzten Zeit vermehrt zum Vergleich der MÜ-Qualität verschiedener Ansätze angewendet (Bentivogli u. a. 2016; Beyer u. a. 2017; Burchardt u. a. 2017). Um ein bestimmtes Phänomen zu untersuchen, werden die Testsuites entweder von Anfang an mit Fokus auf das Zielphänomen selbst konstruiert (Variante 1) oder auf Basis von realen Texten zusammengestellt und anschließend auf das Zielphänomen reduziert (Variante 2) (vgl. King & Falkedal 1990; Koh u. a. 2001). Der Vorteil von Testsuites ist die Möglichkeit der gezielten Untersuchung eines bestimmten Phänomens. Dies geht allerdings bei den beiden Varianten mit einigen Schwierigkeiten einher. Kritisch bei der ersten Variante ist, dass der Text vollständig künstlich konstruiert ist, was mit einem Authentizitätsproblem und Subjektivitätsrisiko verbunden ist (vgl. Balkan u. a. 1994: 34; Koh u. a. 2001). Bei der zweiten Variante liegt die Herausforderung in der Reduzierung der linguistischen Komplexität des Texts zur Untersuchung eines bestimmten Phänomens,

da verschiedene linguistische Phänomene nicht selten in Interaktion stehen (vgl. King & Falkedal 1990). Im Allgemeinen versuchen die Studien bei der Erstellung von Testsuites zwei Aspekte zu realisieren: Abdeckung und Objektivität (vgl. Koh u. a. 2001). Die Abdeckung wird über die Größe der Testsuite erreicht. Die Testsuite von Macketanz u. a. (2018) umfasste mindestens 20 Testsegmente pro Phänomen, um einen ausgewogenen Testsatz zu gewährleisten. Die Objektivität wird durch die Teilnahme von mehreren Linguisten bei der Testsuite-Erstellung optimiert.

Um aus den Vorteilen beider Textressourcen zu profitieren und deren Nachteilen entgegenzuwirken, werden Korpora als Basis für die Erstellung von Testsuites angewendet, eine sog. „Korpusbasierte Testsuite" (vgl. Balkan & Fouvry 1995). Hierbei versucht der Forscher zwar die Sätze so authentisch wie möglich beizubehalten, gleichzeitig eliminiert bzw. reduziert er aber das sog. „noise" (King & Falkedal 1990), sprich linguistische Schwierigkeiten, die für das getestete Problem irrelevant sind, da es nicht zielfördernd ist, unnötig komplexe authentische Sätze zu haben, die den Evaluationsprozess eher erschweren bzw. behindern (vgl. Roturier 2006: 73).

In der vorliegenden Studie wird eine korpusbasierte Testsuite aufgebaut. Zur Erhöhung der Repräsentativität des zugrundeliegenden Korpus umfasst es zehn verschiedene Benutzerhandbücher. Bezüglich der Abdeckung der Testsuite wird jedes Phänomen bzw. jede KS-Regel mit 24 Sätzen repräsentiert. Für eine hohe Objektivität wird bei der Auswahl der Sätze der KS-Checker CLAT[12] (Rösener 2010) verwendet. Zudem werden die Sätze von zwei externen Linguisten zur Qualitätssicherung geprüft. Die detaillierte Beschreibung der korpusbasierten Testsuite der Studie und die Schritte ihrer Erstellung und Qualitätsprüfung sind unter §4.5.3.1 zu finden.

3.4.2.6 Evaluationsteilnehmer

Nach Weber (1998: 65) kann eine Humanevaluation je nach ihrer Zielsetzung von *einsprachigen* oder *zweisprachigen* Teilnehmern durchgeführt werden: Kriterien wie die Verständlichkeit und der Stil des MÜ-Outputs können von *einsprachigen* Teilnehmern bewertet werden. Auf der anderen Seite können die Genauigkeit und Fehlerfreiheit nur mithilfe von *zweisprachigen* Teilnehmern beurteilt werden. (ebd.)

Im Rahmen dieser Studie zeigte die empirische Analyse bei der Bewertung des Stils und der Verständlichkeit der MÜ, dass ein Stil- bzw. Verständlichkeitsfeh-

[12] http://www.iai-sb.de/de/produkte/clat [abgerufen am 23.12.2014]

ler zwar ohne Ausgangstext möglicherweise erkennbar ist, jedoch ist eine nähere Erläuterung des Fehlers, die die erstrebte tiefgründige Analyse ermöglicht, ohne eine Betrachtung des Ausgangstexts erschwert bzw. nicht möglich (mehr dazu unter §4.5.5.2 „Darstellung der Ausgangssätze"). Außerdem beinhaltet die Humanevaluation eine Post-Editing-Aufgabe, die nur von qualifizierten Übersetzern erledigt werden sollte. Vor diesem Hintergrund war der Einsatz von zweisprachigen Teilnehmern bzw. qualifizierten Übersetzern bei der Humanevaluation erforderlich. In dieser Hinsicht steht das Studiendesign in Einklang mit den Erkenntnissen der Post-Editing-Forschung, denen nach bereits Einigkeit darüber herrscht, dass das Post-Editing des MÜ-Outputs nicht von Monolinguisten, sondern von qualifizierten Übersetzern durchgeführt werden sollte (Hansen-Schirra u. a. 2017: 176).

3.4.2.6.1 Einordnung des Studiendesigns

In Anbetracht der verschiedenen dargestellten Methoden lässt sich das Design der vorliegenden Studie wie folgt eingrenzen:

Die *Forscherin als Interessentengruppe* hat im Fokus die *Adäquanz* von fünf Systemen, sprich auf *Makroebene*, vergleichend zu bewerten. Um dies zu realisieren, wird eine *produktorientierte* Evaluation durchgeführt, in der der MÜ-Output der Systeme in einem *Black-Box-Modus* analysiert wird. Im Rahmen der Studie wird eine *deklarative* Evaluation durchgeführt, in der die Qualität des MÜ-Outputs bewertet wird. Parallel wird eine *linguistische* Evaluation durchgeführt, die eine Fehler- und Leistungsanalyse umfasst. Nach einem Mixed-Methods-Ansatz werden die Daten sowohl *quantitativ* als auch *qualitativ* analysiert. Bei den eingesetzten Methoden erfolgt die Bewertung mithilfe von Humanübersetzern und AEMs. Zur Erhöhung des *Objektivitätsgrads* sind mehrere *qualifizierte Humanübersetzer* an der Evaluation beteiligt und es werden zwei *AEMs* genutzt. Als Evaluationsmaterial wird eine *korpusbasierte Testsuite* erstellt und verwendet.

3.4.3 Evaluationsmethoden

Nach der Diskussion des MÜ-Evaluationsdesigns widmet sich dieser Abschnitt den verschiedenen MÜ-Evaluationsansätzen wie auch den Stärken und Schwächen ihrer Methoden.

Zweifellos ist der Bericht des „Automatic Language Processing Advisory Committee" (ALPAC 1966) das bekannteste Ereignis in der Geschichte der MÜ. Trotz der negativen Ergebnisse dieses Berichts, dass die Investition im MÜ-Bereich

nutzlos, zeit- und kostenaufwendig sei, die zur Unterfinanzierung der MÜ-Forschung führten, stellt der Bericht ein wichtiges Dokument über die MÜ-Evaluationsentwicklung dar; dabei ist insbesondere sein Anhang X zu den Evaluationsmethoden von Bedeutung (vgl. Hutchins 1995). Zu den bedeutendsten Forschungen im Bereich MÜ-Evaluation zählen die Forschungsarbeit im Rahmen des jährlichen „Workshop on Statistical Machine Translation“[13] (WMT) sowie der Beitrag der „Defense Advanced Research Projects Agency“ (DARPA) in den 1990ern, die Entwicklung des „Framework for the *Evaluation* of MT“ (FEMTI). In den „Workshops on Statistical Machine Translation“ finden MÜ-Wettbewerbe mit unterschiedlichen Übersetzungs-, Evaluations- und Post-Editingsaufgaben statt. Die DARPA forschte Anfang der 1990er über einen Zeitraum von vier Jahren an dem Thema MÜ, mit dem Ziel, neue MÜ-Ansätze und -methoden im Bereich Evaluation zu entwickeln (White & O'Connell 1996). FEMTI war ein Ergebnis des ISLE-Projekts (International Standards for Language Engineering). Das Ziel von FEMTI war, ein Bild der bis zu diesem Zeitpunkt erfolgten MÜ-Evaluationsarbeit zu schaffen sowie eine Methodik zur Bewertung von MÜ-Systemen unter Berücksichtigung des angestrebten Nutzungskontexts des bewerteten Systems zu etablieren (vgl. King u. a. 2003). Ein weiteres verbreitetes Paradigma zur Bewertung der Übersetzungsqualität ist das MQM-Framework, bei dem die Übersetzungsfehler nach einer umfangreichen Hierarchie von Fehlertypen kategorisiert werden (Lommel u. a. 2013).[14]

Gliederung der MÜ-Evaluationen in Human- und automatischer Evaluation: In der Evaluationsliteratur werden die MÜ-Evaluationsansätze in Human- und automatische Evaluation gegliedert. In der Humanevaluation werden Qualitätskriterien festgelegt und die MÜ von monolingualen Teilnehmern der Zielsprache, bilingualen Teilnehmern bzw. Übersetzern auf Basis dieser Kriterien auf verschiedene Weise bewertet. Die Grundidee einer automatischen Evaluation hingegen ist es, den MÜ-Output mit einer (hochqualitativen) Humanreferenzübersetzung automatisch zu vergleichen. Das Ziel hierbei lautet, herauszufinden, wie ähnlich bzw. unterschiedlich die MÜ im Vergleich zu der Referenzübersetzung ist. Das zeigt, dass eine automatische Evaluation in der Regel nicht rein automatisch erfolgt, denn das Vorhandensein von Humanreferenzübersetzungen ist immer erforderlich. Zudem profitieren die meisten Studien durch eine Methodentriangulation von den Vorteilen der Methoden der beiden Ansätze. Im Folgenden werden die klassischen Methoden der Human- und automatischen Evaluation ausführlich diskutiert. Zudem wird das PE als eine Evaluationstechnik erläutert. Zum

[13] Siehe http://www.statmt.org sowie http://www.statmt.org/wmt19
[14] Siehe www.qt21.eu/mqm-definition

Schluss werden die MÜ-Usability sowie das Eye-Tracking als weitere interdisziplinäre Evaluationsmethoden erörtert.

3.4.3.1 Methoden der Humanevaluation

Wie oben erwähnt, definieren die Studien Qualitätskriterien und bewerten die MÜ gemäß diesen Kriterien. Die Bewertung wird auf Basis von Scoring-Aufgaben, Ranking-Aufgaben, Fehleranalyse, Post-Editing-Faktoren bzw. aufgabenbasiert durchgeführt (vgl. Lavie 2010: 15). Im Folgenden werden die Methoden Scoring, Ranking und Fehleranalyse zusammen mit ihren Vor- und Nachteilen dargestellt. Die Evaluation mithilfe von Post-Editing-Faktoren sowie die aufgabenbasierten Evaluationen im Sinne einer Bewertung der MÜ-Usability sowie durch den Einsatz der Technologie des Eye-Trackings sind primär durch den Einsatz von Bewertern unter den Humanevaluationsmethoden zu klassifizieren. Sie werden jedoch aufgrund ihrer Interdisziplinarität sowie ihres Umfangs gesondert unter §3.4.3.3 und §3.4.3.4 dargestellt.

Scoring ist die herkömmliche Bewertungsmethode, die von DARPA Mitte der 1990er Jahre initiiert wurde (vgl. White & O'Connell 1996; LDC 2002; Lavie 2010: 16). In den Studien von DARPA bewerteten die Evaluatoren die MÜ-Sätze nach den Kriterien „Fluency" und „Adequacy" durch die Vergabe einer Punktzahl auf einer 5-Punkte-Skala (vgl. White & O'Connell 1996). Skalen, wie die Likert-Skala, sind ein weit verbreitetes Bewertungsinstrument im Bereich der MÜ-Evaluation.

Ebenfalls ist die *Ranking*-Methode in der Evaluationsforschung relativ verbreitet. Die Methode wurde im Rahmen einer WMT-Evaluation vorgestellt (Callison-Burch u. a. 2007). Hierbei werden mehrere MÜ-Sätze von der schlechtesten zu der besten gerankt. Dementsprechend – anders als die Scoring-Methode – werden die Sätze beim Ranking in einem direkten Vergleich relativ zueinander bewertet.

Ein Vergleich der beiden Methoden Scoring vs. Ranking zeigte (vgl. ebd.), dass Scoring die Arbeit der Bewerter erschwerte, da sie keine genaue Anweisung hatten, wie sie ihre Bewertung und etwaige Fehler quantifizieren sollen. Dementsprechend legte jeder Bewerter seine eigenen Faustregeln zum Scoring fest oder verwendete die Skala für ein relatives – anstatt absolutes – Scoring. Dies wiederum führte dazu, dass das Ranking ein höheres Inter-Annotator- und Intra-Annotator-Agreement als das Scoring zeigte. Zudem fanden Callison-Burch et al. (ebd.) es vorteilhaft, dass das Ranking eine differenzierte Unterscheidung zwischen den Übersetzungen ermöglicht, die beim Scoring nicht möglich wäre.

Trotz der Vorteile des Rankings arbeiten viele Studien mit Scoring. Die Teilnehmer der WMT-Evaluationen berichten von mehreren Fällen, in denen das Ranking von Übersetzungen besonders schwierig war (Denkowski & Lavie 2010).

Erstens müssen sich die Bewerter im Fall von längeren Sätzen diese Sätze gut merken. Dafür hat jeder Bewerter seine eigene Strategie für den Umgang mit langen Sätzen. Jedoch kann es je nach Satzlänge und -komplexität schwer sein, konsequent eine bestimmte Bewertungsstrategie beizubehalten. Dies führt zu einem verringerten Intra-Annotator-Agreement. Eine weitere Schwierigkeit beim Ranking tritt bei der Bewertung von mehreren qualitativ ähnlichen Übersetzungen auf, die allerdings unterschiedliche Fehlertypen beinhalten (z. B. einen Rechtschreibfehler vs. Großschreibfehler). Hier besteht die Möglichkeit, dass der Bewerter beide Fehler für vergleichbar hält und der MÜ das gleiche Qualitätsniveau zuweisen würde, was im Normalfall der Idee eins Rankings widerspricht. Diese Schwierigkeit wird in der Regel umgangen, indem die Studie die Vergabe von identischen Bewertungen, sogenannten „ties", zulässt; somit hat der Bewerter die Möglichkeit, nach seiner Bewertung ähnlich problematische Fehler identisch zu gewichten. Die Verwendung von „ties" bringt allerdings das Problem mit sich, dass die Ergebnisse der Studie möglicherweise viele identische Bewertungen beinhalten, was im Endeffekt einen Vergleich der Übersetzungen unmöglich macht. Ein weiteres Problem des Ranking ist, dass der Schwierigkeitsgrad des Fehlers schwer messbar sein kann (Costa u. a. 2015), z. B. sollten zwei Gruppen bestehend aus je fünf MÜ-Sätzen gerankt werden: Die erste Gruppe beinhaltet mehrere richtige MÜ-Sätze und einen falschen MÜ-Satz mit einem einfachen Fehler; die zweite Gruppe beinhaltet mehrere falsche MÜ-Sätze mit schwerwiegenden Fehlern. Sowohl der MÜ-Satz mit dem einfachen Fehler in der ersten Gruppe als auch der schlechteste MÜ-Satz mit dem schwerwiegenden Fehler aus der zweiten Gruppe werden trotz des unterschiedlichen Schwierigkeitsgrads der aufgetretenen Fehler auf Platz 5 gerankt. Dementsprechend muss für die Auswertung der Fehlerschwierigkeitsgrade das Ranking um andere Bewertungstechniken erweitert werden.

Angesichts der Vor- und Nachteile der beiden Methoden kann nur bei jeder Studie individuell auf Basis der analysierten Sätze und Fragestellungen entschieden werden, welche der beiden Methoden sich eignet. Sollten relativ lange Sätze oder eine große Anzahl von Sätzen verglichen werden, wäre es ratsam, das Scoring zu verwenden. Dabei muss darauf geachtet werden, dass die Bewerter genaue Anweisungen für das Scoring erhalten. Geht es primär darum, Übersetzungen im Verhältnis zueinander zu bewerten, käme eher das Ranking in Frage. Auch hier sollte die Eignung der Fehlertypen für das Ranking geprüft werden, z. B. ist das Ranking von ähnlichen Fehlertypen in der Regel für viele Bewerter problematisch. In der vorliegenden Studie wurde das Scoring im Rahmen der Humanevaluation zur Bewertung der Qualität der MÜ in den Untersuchungsszenarien vor und nach dem Einsatz der KS auf einer 5-Punkte-Likert-Skala ange-

wendet. Die Gründe für die Verwendung des Scorings sind unter §4.5.5.2 „Form der Evaluationsfragen“ dargestellt.

Das Scoring und Ranking ermöglichen zwar den Vergleich des Outputs mehrerer MÜ-Systeme, sie geben jedoch keinen Aufschluss über den Hintergrund der erhaltenen Punktzahl. Dies ist anhand einer weiteren bewährten Methode der Humanevaluation, der *Fehleranalyse* (auch *Fehlerannotation* genannt) des MÜ-Outputs, möglich. Sie basiert auf der herkömmlichen Idee, von Fehlern zu lernen, um Verbesserungen zu erzielen. In einer Fehleranalyse wird ein Fehlerschema festgelegt, in dem die Fehler in Kategorien klassifiziert sind. Anschließend wird die MÜ entsprechend dieses Schemas annotiert. Die Fehlerannotation kann primär auf Basis des Ausgangstexts (sog. Free annotation) oder durch den Vergleich mit einer oder mehreren Referenzübersetzungen (sog. Flexible reference-based annotation) erfolgen (Fishel u. a. 2012). Bei der ersten Annotationsstrategie (Free annotation) kann das Risiko eines Punktabzugs bei einer korrekten MÜ aufgrund ihrer Dissimilarität zur Referenzübersetzung vermieden werden (ebd.). Auf der anderen Seite vereinfacht die zweite Annotationsstrategie (Flexible reference-based annotation) dem Annotator die Annotationsaufgabe und erhöht das Inter-Annotator-Agreement (ebd.).

Nach der Annotation werden die Ergebnisse auf verschiedene Weise analysiert. Hutchins & Somers (1992: 164) finden, dass „the most useful practical information is obtained from error counting“. Gleichzeitig räumen sie ein (ebd.), dass „for many purposes, however, the simple counting of errors is insufficient. What is needed is a classification of errors by types of linguistic phenomenon and by relative difficulty of correction“. Die Ermittlung der Fehleranzahl gibt Aufschluss über die Qualität des Outputs eines bestimmten Systems. Vergleicht man mehrere Systeme, ist es für die Objektivität des Vergleiches essentiell, neben der Fehleranzahl die Fehlertypen zu berücksichtigen, denn *ein* orthografischer Fehler (z. B. Kommasetzung) ist in Hinsicht auf den Korrekturaufwand mit *einem* syntaktischen Fehler (z. B. Wortstellung) nicht vergleichbar. Von diesen Ergebnissen profitieren mehrere Zielgruppen (Stymne 2013): Die Entwickler und Forscher identifizieren die häufigsten und schwerwiegendsten Fehler und können sich gezielt darauf konzentrieren. Die Benutzer können ebenfalls die gängigen Fehlertypen erkennen und somit den System-Output besser verstehen. Auch die Firmen haben bei der Anschaffung eines Systems die Möglichkeit, die Stärken und Schwächen mehrerer Systeme zu vergleichen.

Es wurden zahlreiche Fehlertaxonomien entwickelt, die je nach Bewertungssetting in ihrer Granularität variieren. Correa (2003) entwickelte eine Fehlertaxonomie für die Evaluation von MÜ in einem Glass-Box-Kontext, die eine relativ hohe Granularität aufwies. Sie beinhaltete die Kategorien Qualität-Score,

Input-Segmentfehler, Segmentierungsfehler, Markup-Fehler, unbekanntes Wort, Fehler bei benannter Entität, Fehler bei der Analyse des Ausgangstexts, lexikalischer Fehler im Zieltext, Grammatikfehler im Zieltext und stilistischer Fehler im Zieltext. Eine weitere bekannte Taxonomie ist die von Flanagan (1994), die zum Vergleich von konkurrierenden Systemen mithilfe von Endnutzern entwickelt wurde. Die Taxonomie wies eine hohe Granularität auf und berücksichtigte die sprachlichen Unterschiede in zwei Sprachenpaaren (ebd.). Sie bestand für das Sprachenpaar Englisch-Französisch aus den Kategorien Rechtschreibfehler, im Wörterbuch nicht vorhandenes Wort, inkorrekter Akzent, Großschreibungsfehler, Elision, inkorrekte Verbflexion, inkorrekte Nominalflexion, weitere inkorrekte Flexion, Wortstellungsfehler, inkorrekte Wortkategorie, inkorrektes/fehlendes Pronomen, inkorrekter/fehlender Artikel, inkorrekte/fehlende Präposition, negative/falsch platzierte/fehlende Partikel, inkorrekte Konjunktion, Kongruenzfehler, Fehler bei der Identifizierung der Satzgrenze und inkorrekte Wortauswahl. Für das Sprachenpaar Englisch-Deutsch wurde die Fehlertaxonomie um die Kategorien falsches/fehlendes Relativpronomen, inkorrekte Kasusendung und inkorrekte Zeichensetzung erweitert. Die Fehlerkategorien wurden vom Nutzer je nach seinen Kriterien gerankt. Sprachenpaarspezifische Fehler spielen bei der Fehleranalyse eine wesentliche Rolle. Condon u. a. (2010) identifizierten in ihrer SMÜ-Fehleranalyse eine Reihe von Sprachenpaarspezifischen Fehlern, die nicht aufgrund fehlender Beispiele in den Trainingsdaten, sondern aufgrund sprachlicher Unterschiede zwischen den analysierten Sprachen auftreten. Llitjós u. a. (2005) entwarfen eine zweistufige Fehlertypologie mit dem Ziel, die MÜ nach dieser Typologie zu posteditieren und die Korrekturen automatisch in die Übersetzungsgrammatikregeln und die lexikalischen Einträge zu übernehmen. Die Hauptkategorien der Typologie waren fehlendes Wort, extra eingefügtes Wort, Wortstellungsfehler, inkorrektes Wort und Kongruenzfehler.

Wie die obige Darstellung zeigt, weisen die Fehlertypologien viele Gemeinsamkeiten auf. Dementsprechend zielten Vilar u. a. (2006) mit ihrer Fehlertaxonomie darauf ab, eine explizite Fehlerklassifikation zu entwickeln. Sie wurde in Anlehnung an die Fehlertypologie von Llitjós u. a. (2005) entworfen und besteht aus einer dreistufigen Taxonomie, die fünf Hauptkategorien (fehlende Wörter, Wortstellungsfehler, falsche Wörter, unbekannte Wörter und Zeichensetzungsfehler) beinhaltet. Die Fehlertaxonomie von Vilar u. a. (2006) wurde häufig in der Evaluationsforschung als Grundlage verwendet (u. a. in Avramidis & Koehn 2008; Bojar 2011; Popović & Burchardt 2011). Eine weitere aktuelle Fehlerhierarchie, die häufig angewandt wird, ist das Multidimensional Quality Metrics (MQM) Framework (Lommel u. a. 2013). Es handelt sich dabei nicht um eine Qualitätsmetrik, sondern um ein Framework zur Erstellung von Metriken. Ausgehend

von der Annahme, dass kein einzelnes festes Fehlerkategorisierungsschema zur Bewertung der Qualität einer Vielzahl unterschiedlicher Übersetzungsprojekte verwendet werden kann, stellt MQM eine flexible Fehlerhierarchie von über 100 Fehlertypen zur Verfügung, mit dem Ziel, daraus maßgeschneidert die relevante Teilmenge von Fehlertypen auszuwählen (Lommel u. a. 2013). Als Basis für die Hierarchie dienen die Hauptdimensionen „accuracy, fluency, design, verify und internationalization"; darunter folgen die weiteren Ebenen der Fehlertypen, aus denen der Bewerter je nach Projektanforderungen und -erwartungen auswählen kann (ebd.).

Neben der Entwicklung der Fehlertypologien beschäftigten sich die Forscher im Bereich der Fehleranalyse mit ihrer Modalität d. h. der Durchführung von *manuellen* Fehlerannotationen (meist mithilfe von Annotationstools, wie BLAST[15] (Stymne 2011)) und *automatischen* Fehlerannotationen. In Studien wie Fishel u. a. (2012) und Elliott u. a. (2004) wurden die MÜ-Fehler manuell annotiert. Mithilfe der manuell annotierten MÜ-Korpora bewerteten Fishel u. a. (2012) zwei automatische Annotationstools zur MÜ-Diagnostik und -Bewertung: Addicter (Zeman u. a. 2011) und Hjerson (Popović 2011; Popović & Ney 2011). Automatische Annotationstools wurden zur Beschleunigung der Fehlerannotation und Reduzierung ihrer Kosten entwickelt. Das Arbeitsprinzip von diesen automatischen Annotationstools ist vergleichbar mit dem der AEMs (siehe §3.4.3.2). Bei Addicter (Zeman u. a. 2011) werden die MÜ und die Referenzübersetzungen aligniert, um die auftretenden Fehlertypen zu ermitteln. Bei Hjerson (Popović & Ney 2011) werden – zur Realisierung desselben Ziels – die WER- und PER-Metrik auf Wortebene in der MÜ zerlegt.

Die in der vorliegenden Studie angewandte Fehlertaxonomie wurde nach einem Bottom-up-Ansatz ausgehend von Vilar u. a. (2006) festgelegt. Die genaue Vorgehensweise für die Erstellung der Studientaxonomie ist unter §4.5.4.1 dargestellt. Die Fehlerannotation wurde zur Untersuchung des Einflusses des KS-Einsatzes in Bezug auf die Fehleranzahl und die Fehlertypen vor und nach der Anwendung von KS-Regeln verwendet. Die Annotation erfolgte lediglich auf Basis des Ausgangstexts (Strategie der „Free Annotation" (Fishel u. a. 2012)), um das Risiko eines Punktabzugs bei korrekter MÜ, die einer Referenzübersetzung unähnlich ist, auszuschließen. Die Annotation wurde manuell unter einer Unterscheidung zwischen Fehlern innerhalb und außerhalb der KS-Stelle (siehe §4.5.2.1) durchgeführt. Diese Unterscheidung ist maßgebend für die Studie und im Rahmen einer automatischen Fehlerannotation komplex programmierbar.

[15] BLAST (the BiLingual Annotator/Annotation/Analysis Support Tool) ist ein Open-Source-Tool zur Fehlerannotation vom MÜ-Ouput (Stymne 2011), das u. a. von Flanagan (1994) und Vilar u. a. (2006) verwendet wurde.

Die Methoden der Humanevaluation finden vielfach Anwendung. Sie ermöglichen es, Qualitätskriterien und Übersetzungsfehler quantitativ zu bewerten. Zudem können die Daten durch eine Methodentriangulation, z. B. von Scoring und Fehlerannotation, wie in der vorliegenden Studie, qualitativ analysiert werden. Auf der anderen Seite werden die Humanevaluationsmethoden wegen des damit verbundenen Zeit- und Kostenaufwands sowie der Subjektivität einer Humanbeurteilung kritisiert. Eine mögliche Folge der Subjektivität ist eine niedrige Konsistenz der Ergebnisse im Sinne eines niedrigen Interrater- and Intrarater-Agreement-Niveaus (vgl. Lavie 2010: 14; Doherty 2012: 44). Dennoch gibt es mehrere Möglichkeiten, der Subjektivität entgegenzuwirken, u. a. die Verwendung einer konsistenten und systematischen Fehlertypologie bei einer Fehleranalyse (Flanagan 1994) und klar definierter Qualitätskriterien beim Scoring (vgl. Doherty 2017). Um der Kostenproblematik der Humanevaluation gegenzusteuern wird in den letzten Jahren das Crowdsourcing vermehrt angewendet.

Crowdsourcing ist das „concept of outsourcing a job to an undefined group of people over the internet" (Gerlach 2015: 109). Eine der häufig verwendeten Plattformen für Crowdsourcing in der Forschungsarbeit ist der „Amazon Turk Mechanical". Sie wird in den Bereichen MÜ, Post-Editing (Mitchell u. a. 2014) und NLP (Natural Language Processing) (Snow u. a. 2008) verwendet. Studien wie „Cheap and fast – but is it good?" (ebd.) verglichen die Daten von Experten-Annotatoren und Nicht-Experten von Crowdsourcing und kamen zu dem Ergebnis, dass eine kleine Anzahl von Crowdsourcing-Teilnehmern (ca. 4 Personen) Ergebnisse liefern kann, die mit denen eines Experten-Annotatoren vergleichbar sind. Außerdem zeigte die Studie eine Möglichkeit auf, Verzerrungen einzelner Teilnehmer zu modellieren und zu korrigieren, wenn sowohl Experten- und Nicht-Experten-Daten verfügbar sind (ebd.). Ebenfalls fanden Mitchell u. a. (2014) heraus, dass die Evaluationsergebnisse von Domain-Spezialisten und Crowdsourcing-Teilnehmern ähnlich sind und eine starke positive Korrelation aufweisen.

Der Vorteil des Crowdsourcing liegt darin, dass die Evaluationen mit einer großen Zahl an Teilnehmern kostengünstig und in kurzer Zeit durchgeführt werden können (Schenk & Guittard 2011; Liu u. a. 2012). Gleichzeitig bringt die Methode die Gefahr mit sich, dass die Qualität der gelieferten Arbeit je nach Schwierigkeitsgrad der Aufgaben variiert und u. U. leiden kann (vgl. Schenk & Guittard 2011; Liu u. a. 2012; Stevens 2018). Crowdsourcing ist daher primär für einfache Aufgaben geeignet (Stevens 2018). Außerdem ist eine Interaktion mit den Teilnehmern erschwert (Liu u. a. 2012).

In der vorliegenden Studie fiel die Entscheidung gegen die Verwendung von Crowdsourcing. Zur Gewährleistung der gelieferten Qualität legte die Forscherin Wert auf eine sorgfältige Auswahl und den Nachweis der Bewerterprofile.

Zudem bestand die Evaluation aus einer relativ komplexen Aufgabe, in der die Übersetzungsqualität auf zwei Ebenen und differenziert nach fünf Qualitätskriterien bewertet werden musste. Hierfür waren neben den Sprachkompetenzen fundierte Translationskompetenzen erforderlich. Die Testphase erstreckte sich über 3 bis 4 Wochen (durchschnittlich 2 Tests täglich pro Teilnehmer). Nach dem Testablauf (siehe §4.5.5.2 „Ablauf der Evaluation“) erhielt die Forscherin die beantworteten Tests täglich, prüfte sie auf Vollständigkeit und kontaktierte den Teilnehmer bei Bedarf, um fehlende Inputs zu vervollständigen. Im Anschluss erhielt der Teilnehmer die Tests für den kommenden Tag. Dieser Ablauf sorgte für Qualitätssicherung der Ergebnisse und wäre im Falle eines Crowdsourcing nicht realisierbar. Außerdem war – wie Abschnitt §4.5.5.3 zeigt – eine Durchführung mit acht Teilnehmern ausreichend, da ab dem sechsten Teilnehmer die akkumulierten Qualitätswerte begannen, sich zu stabilisieren. Dementsprechend war ein Crowdsourcing nicht erforderlich.

3.4.3.2 Automatische Evaluation: automatische Evaluationsmetriken

Mit der zunehmenden Entwicklung und stetigen Weiterentwicklung von MÜ-Systemen entstand der Bedarf nach einer zeit- und kosteneffizienten Evaluationsmethode. Während eines Optimierungsprozesses benötigt der Systemanbieter eine schnelle quantifizierbare Bewertung, die ihm einen Indikator zu den vorgenommenen Anpassungen liefert (vgl. Doherty 2012: 44). Das war der Anreiz für die Entwicklung automatischer Evaluationsmetriken (AEMs) (ebd.). AEMs sind „scripts or software programs that perform Translation Quality Assessment (TQA) using an explicit and formalized set of linguistic criteria to evaluate the MT output, typically against a human reference translation or ‚gold standard'“ (Doherty 2017: 4). Aus dieser Definition lässt sich die Grundidee der automatischen Qualitätsevaluation anhand der AEMs ableiten, nämlich den MÜ-Output mit einer oder mehreren hochqualitativen Referenzübersetzungen (sog. Goldstandard-Humanübersetzung) zu vergleichen, um zu messen, wie ähnlich bzw. unterschiedlich der MÜ-Output im Vergleich zu der Referenzübersetzung ist.

Es existieren zahlreiche automatische Evaluationsmetriken (AEMs), deren Entwicklung auf die 90er Jahre zurückgeht. Bereits 1992 bewerteten Su u. a. (1992) die Übersetzungsqualität nach dem Edit-Distance-Prinzip von Levenshtein (1966). Sie definierten die Edit-Distance in Bezug auf die „editing efforts needed to edit the raw translation output into the revised version“ und ermittelten den Bearbeitungsaufwand anhand der Anzahl der „Key strokes“ bei vier Bearbeitungsoperationen: „insertion, deletion, replacement and swap“ (ebd.: 435). Danach entwickelten Nießen u. a. (2000) die Metrik WER (Word Error Rate) ebenfalls auf

Basis des Edit-Distance-Prinzips von Levenshtein (1966: 707) gemessen als die „number of insertions, deletions and substitutions between the produced translation and one predefined reference translation". Eine Schwäche der AEMs, die schon in den Scores von WER zu beobachten war, ist der Umgang mit unterschiedlicher Wortreihenfolge im MÜ-Output im Vergleich zu der Referenzübersetzung (vgl. Han u. a. 2017). Eine abweichende Wortreihenfolge deutet nicht zwangsläufig auf eine fehlerhafte Übersetzung hin. Sie kann sich zwar durch einen Wortstellungsfehler ergeben oder aber aufgrund einer weiteren möglichen (korrekten) Übersetzung erfolgen (vgl. ebd.). In WER war der Score sehr niedrig, wenn die Wortreihenfolge im MÜ-Output von der der Referenzübersetzung abwich (vgl. ebd.).

Vor diesem Hintergrund berücksichtigten Snover u. a. (2006) bei ihrer Metrik TER (Translation Edit Rate), einer weiteren Edit-Distance-Metrik, die Wortstellungsproblematik:

> TER adds block movement (jumping action) as an editing step. The shift option performs on a contiguous sequence of words within the output sentence. […] For the edits, the cost of the block movement, any number of continuous words and any distance, is equal to that of the single word operation, such as insertion, deletion and substitution. (Han u. a. 2017: 5)

Andere AEMs arbeiten nach den Informationsabrufkonzepten (information retrieval concepts) der Genauigkeit und Trefferquote[16] mit einzelnen Wörtern (unigrams) oder mit längeren n-grams (Doherty 2017). Darunter fällt z. B. die weit verbreitete Bewertungsmetrik BLEU (Papineni u. a. 2002). Bei der Evaluation mit BLEU (BiLingual Evaluation Understudy) wird gemessen, wie viele Wörter sich im Übersetzungsoutput mit der Referenzübersetzung überschneiden, wobei Wortfolgen höher bewertet werden (ebd.). Auf diese Weise geht BLEU mit der Wortstellungsproblematik um. Sollten mehrere Referenzübersetzungen vorhanden sein, so wird in BLEU die Referenzübersetzung mit der ähnlichsten Länge zu dem Übersetzungsoutput bei der Evaluation herangezogen (Han u. a. 2017). In BLEU werden die Scores anschließend über den gesamten Korpus gemittelt, um eine Abschätzung der Gesamtqualität der Übersetzung zu erhalten (Papineni u. a. 2002).

[16] Die Genauigkeit (precision) bezieht sich auf den Anteil der Wörter im MÜ-Output, der auf Basis der Referenzübersetzung korrekt übersetzt wurde; die Trefferquote (recall) ist der Anteil der gemäß der Referenzübersetzung korrekten Wörter, der im MÜ-Output wiedergegeben wurde. Sollte beispielsweise eine Referenzübersetzung insgesamt 10 Wörter umfassen, während der MÜ-Output aus 8 Wörtern besteht, die alle korrekt übersetzt wurden, würde man hier von einer 100%-en Genauigkeit und einer 80%-en Trefferquote sprechen. (Doherty 2017)

Es folgten weitere bekannte AEMs, wie z. B. NIST (Doddington 2002), ROUGE (Lin & Hovy 2003), METEOR (Banerjee & Lavie 2005), ATEC (Wong & yu Kit 2009), PORT (Chen u. a. 2012), LEPOR (Han u. a. 2012) und hLEPOR (Han u. a. 2013). Im Prinzip versuchten die neuentwickelten AEMs die Schwächen der vorangegangenen Metriken zu behandeln.

Mit der automatischen Evaluation lassen sich die MÜ-Systeme im Vergleich zu der Humanevaluation zeit- und kosteneffizienter bewerten. Daher wird in der Regel die Effektivität der AEMs durch ihre Korrelationen mit den Ergebnissen von Humanevaluationen beurteilt. BLEU wurde aufgrund der hohen Korrelation ihrer Ergebnisse mit denen der Humanevaluation in zahlreichen Studien sowie mehrmals in den ACL-Workshops on Statistical Machine Translation als einzige Evaluationsmetrik angewandt (vgl. Callison-Burch u. a. 2007). Mit der Entwicklung der neuen AEM-Generation verglich der ACL-Workshop 2009 verschiedene Metriken, mit dem Ziel, die Korrelation ihrer Ergebnisse mit denen der Humanevaluation zu messen. Die Untersuchung zeigte, dass mehrere Metriken der neuen Generation BLEU hinsichtlich der Korrelation mit der Humanevaluation übertreffen (Callison-Burch u. a. 2009).

Im Vergleich zur Humanevaluation wird die automatische Evaluation ohne bilinguale Bewerter oder Muttersprachler durchgeführt. Daher bietet die automatische Evaluation den Systemanbietern eine flexible, schnelle und kosteneffiziente Lösung zur Bewertung von (fortlaufenden) Systemoptimierungen. Gleichzeitig werden zur Anfertigung der Referenzübersetzung Übersetzer gebraucht. Das Argument, dass die automatische Evaluation eine objektivere Methode im Vergleich zur Humanevaluation sei, ist daher fraglich (vgl. Doherty 2017): Eine automatische Evaluation basiert auf einer Referenz*human*übersetzung und wird anhand eines von *Menschen* entwickelten Programms ausgeführt. Die Referenzübersetzung ist ein Output, der auf einer Reihe von subjektiven Entscheidungen seines Humanübersetzers basiert. Ebenfalls wird der Aufbau des AEM-Programms von seinem Entwickler subjektiv festgelegt. Dennoch lässt sich die Objektivität der Evaluation durch die Verwendung von mehreren Referenzübersetzungen und AEMs steigern. Ferner stoßen die *automatischen* Metriken oft auf ähnliche Schwierigkeiten wie die der *automatischen* Übersetzung, nämlich bei der Erkennung bzw. Bewertung von syntaktischen und semantischen Äquivalenzen. Es wurden zwar mehrere AEMs entwickelt, die tiefere linguistische Analysen ausführen (vgl. Giménez & Màrquez 2008; Padó u. a. 2009; Liu u. a. 2011), dennoch bleibt die herkömmliche syntaktische und semantische Komplexität eine Hürde für die Automatisierung, sei es bei der MÜ oder bei der Evaluation. Daher wird bei der Bewertung häufig die automatische Evaluation mit einer Humanevaluation kombiniert. Durch die Humanevaluation kann nicht nur die syntaktische und

semantische Komplexität bewältigt werden, auch eine Interpretation der rein numerischen Ergebnisse der AEMs wird möglich.

In dieser Studie wurden eine Humanevaluation und eine automatische Evaluation kombiniert durchgeführt. Die angewandten AEMs waren TERbase (Snover u. a. 2006; Gonzàlez & Giménez 2014: 19) und hLEPOR (Han u. a. 2013). Die Auswahl der beiden AEMs wird in §4.5.6.3 begründet.

3.4.3.3 Post-Editing als MÜ-Evaluationstechnik

Eine der wichtigen Forschungsfragen, mit der sich die Post-Editing-Forscher beschäftigen, ist „to what extent MT output texts are acceptable, and how much human effort is necessary to improve such imperfect texts“ (Allen 2003: 298). Dementsprechend ist das Post-Editing (PE) als eine hilfreiche Evaluationstechnik der MÜ zu betrachten, die aufgrund der Teilnahme von Post-Editoren und Bewertern unter die Humanevaluationsmethoden fällt. Die zahlreichen Korpora von post-editierten MÜ (z. B. Callison-Burch u. a. 2012; Potet u. a. 2012), die mit der Entwicklung der MÜ und des PE erstellt wurden, wurden hauptsächlich zur Evaluation der MÜ-Qualität eingesetzt (Wisniewski u. a. 2013). Bevor genauer auf die MÜ-Evaluation anhand des Post-Editing eingegangen wird, werden im Folgenden die Definition, Arten und Tools des PE kurz erläutert.

Das PE wird als die „correction of raw machine translated output by a human translator according to specific guidelines and quality criteria“ (O'Brien 2011: 197f.) definiert. Somit ist das PE ein Verfahren der Nachbearbeitung eines maschinell übersetzten Texts, das ein erforderliches Qualitätsniveau des Texts ermöglicht. Das PE sollte von qualifizierten Humanübersetzern, nicht von Laien und keinesfalls von monolingualen Bewertern, durchgeführt werden, da sonst die Gefahr inhaltlicher Fehler nicht ausgeschlossen werden kann (Hansen-Schirra u. a. 2017: 176).

Man unterscheidet im Allgemeinen zwischen zwei Arten des PE, „Light Post-Editing“ und „Full Post-Editing“ (vgl. O'Brien 2010a; O'Brien 2010b): Ein „Light Post-Editing“ ist eine schnelle Bearbeitung und Korrektur der wesentlichen Fehler mit dem Fokus, den Hauptinhalt in einer verständlichen und genauen Form bereitzustellen; ein „Full Post-Editing“ ist hingegen eine umfassende Korrektur zur Qualitätssteigerung. Bei Letzterem wird versucht, den MÜ-Output in eine qualitativ vergleichbare Humanübersetzung umzuwandeln (Wagner 1987). Je nach dem erforderlichen Qualitätsniveau kann die entsprechende PE-Art bei der Evaluation umgesetzt werden.

Zur Unterstützung solcher Evaluationen werden in der Post-Editing-Forschung Tools wie TranslogII (Carl 2012), CASMACAT (Alabau u. a. 2013) und PET[17] (Aziz u. a. 2012) verwendet, mit deren Hilfe der PE-Prozess erfasst und untersucht werden kann. Dies wiederum ermöglicht die Ermittlung der MÜ-Probleme und somit die Bewertung ihrer Qualität.

Sowohl in der Praxis als auch in der Forschung wird intensiv untersucht, ob und in welchen Szenarien eine MÜ kombiniert mit einem PE effizienter als eine Humanübersetzung sein kann (Drugan 2013: 99). Um diese Frage zu beantworten, wird auf unterschiedliche Weise versucht, den Aufwand des PE zu quantifizieren. In dieser Hinsicht definierte Krings (2001) drei Dimensionen des PE-Aufwands – der zeitliche, technische und kognitive Aufwand des PE: Mit dem *Zeitaufwand* misst man die Zeit, die das PE in Anspruch nimmt. Im *technischen Aufwand* werden die Bearbeitungsvorgänge, die zur Erstellung der post-editierten Version erforderlich sind, berechnet (i. d. R. in Bezug auf die Anzahl der Tastaturanschläge beim Einfügen und Löschen von Zeichen). Bei dem *kognitiven Aufwand* werden die Schritte der Identifizierung der MÜ-Fehler und Durchführung der erforderlichen Korrektur genauer untersucht. (ebd.)

Während der zeitliche und der technische Aufwand sich in Dauer und Anzahl der Modifikationen typischerweise messen lassen (vgl. Tatsumi 2009; Tatsumi & Roturier 2010; Specia 2011), wird der kognitive Aufwand nach unterschiedlichen Vorgehensweisen gemessen. Frühere Studien maßen den kognitiven Aufwand anhand von Qualität-Scores, die auf dem von Humanbewertern wahrgenommenen PE-Aufwand basieren (Specia u. a. 2010; Specia 2011; Koponen 2012; Popović u. a. 2014). Weitere, aktuellere Studien setzen die Technologie des Eyetracking zur Untersuchung des kognitiven PE-Aufwands ein (vgl. O'Brien 2011; Carl u. a. 2015; Nitzke 2019) (siehe §3.4.3.4). Grundsätzlich „hat der kognitive Aufwand natürlich mit Tastatureingaben und Zeit zu tun" (Hansen-Schirra u. a. 2017: 181). Die Relationen zwischen diesen PE-Parametern wurden in mehreren Studien im Kontext der SMÜ (Popović u. a. 2014), der phrasenbasierten SMÜ (Koponen 2012) und in einer aktuellen Studie für SMÜ im Vergleich zu NMÜ (Toral u. a. 2018) untersucht.

Die aktuellen Evaluationsstudien zeigen, dass der NMÜ-Output deutlich flüssiger ist und weniger Morphologie- und Grammatikfehler im Vergleich zum davor dominanten Ansatz, PBMÜ, aufweist (vgl. Bentivogli u. a. 2016; Klubička u. a. 2017; Toral & Sanchez-Cartagena 2017). Diese auffällige Flüssigkeit verbirgt in

[17] PET ist online unter http://www.clg.wlv.ac.uk/projects/PET verfügbar. Weitere Tools und Datensätze für posteditierte MÜ sind unter https://staffwww.dcs.shef.ac.uk/people/L.Specia/resources.html abrufbar.

manchen Fällen jedoch Genauigkeits- und Stilistikfehler (vgl. Volk 2018; Vardaro u. a. 2019), sodass eine entwicklungsgemäße Anpassung der Details der PE-Aufteilung in „light-“ und „full-PE“ in naher Zukunft zu erwarten ist. Volk (2018: 8) beschreibt das PE bei der NMÜ wie folgt: “[I]t is also acknowledged that this requires even more attention and a higher cognitive workload for professional translators who are post-editing MT output. The errors of the MT system are now even more difficult to spot.“ Im Abschnitt §3.5 werden Evaluationsstudien im Kontext der KS, die mithilfe der PE-Technik durchgeführt wurden, dargestellt.

Im Rahmen der Humanevaluation der vorliegenden Studie führten die Teilnehmer ein Light-PE durch. Vergleichbar mit der Vorgehensweise von Snover u. a. (2006) wurde die posteditierte MÜ als Referenzübersetzung bei der automatischen Evaluation verwendet (Mehr dazu unter §4.5.6.1). Ferner beleuchteten die posteditierten Übersetzungen die MÜ-Fehler und ermöglichten eine qualitative Analyse im Zusammenhang mit den Qualitätskriterien.

3.4.3.4 Weitere (interdisziplinäre) MÜ-Evaluationsmethoden

3.4.3.4.1 Evaluation der MÜ-Usability

In einer weiteren Technik der MÜ-Evaluation wird die Usability des MÜ-Outputs mit dem Endnutzer im Fokus bewertet. Wie kann man die MÜ anhand des Usability-Konzepts bewerten? Um diese Frage beantworten zu können, wäre es zunächst hilfreich, die Definition der Usability näher zu betrachten. Die ISO (2002) definiert die Usability als „the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use“. Entsprechend dieser Definition war das allgemeine Ziel der MÜ-Evaluationsstudien, herauszufinden, inwiefern die Nutzer der MÜ-Systeme ihre Ziele vom MÜ-Output als Produkt auf effektive und effiziente Weise realisieren können und wie zufrieden sie mit diesem Produkt sind.

Eine der ersten Studien im Bereich der Usability des MÜ-Outputs ist von Tomita u. a. (1993), in der sie die Verständlichkeit des MÜ-Outputs mehrerer MÜ-Systeme untersuchten und verglichen. Der Untersuchungsgegenstand waren Texte zum Leseverstehen der TOEFL-Prüfung,[18] die mit den MÜ-Systemen ins Japanische übersetzt wurden und deren Fragen von Japanischmuttersprachlern beantwortet wurden. Die Idee dabei war, dass die Testscores zeigen, inwiefern die MÜ-Systeme in der Lage sind, den Probanden die semantischen Inhalte der Texte zum Leseverstehen zu vermitteln. (ebd.) In einer weiteren Studie verglichen

[18] TOEFL (Test of English as a Foreign Language) ist ein weltweit anerkannter Englisch-Sprachtest.

Doherty & O'Brien (2012) die Usability der rohen MÜ einer technischen Hilfedokumentation zum Thema Online-Dateispeicherung in den Sprachen Spanisch, Französisch, Deutsch und Japanisch mit der der englischen Originaldokumentation. Basierend auf der Definition der Usability und mithilfe von potenziellen Endnutzern maßen Doherty und O'Brien (ebd.), ob die potenziellen Endnutzer in der Lage wären, die technischen Aufgaben anhand der Dokumentationsanweisungen erfolgreich abzuschließen (goal completion), die Dauer der Aufgabenlösung (effectiveness), die Anzahl der erfolgreich abgeschlossenen Aufgaben unter Berücksichtigung der Gesamtdauer (efficiency) sowie die Benutzerzufriedenheit mit den Anweisungen. Die Studie wurde mithilfe der Eye-Tracking-Technologie durchgeführt (mehr zum Eye-Tracking folgt im nächsten Abschnitt). Shubert u. a. (1995) verwendeten die Usability-Technik zur Untersuchung von mit Simplified English (SE) verfasster Verfahrensdokumentation im Bereich der Flugzeugwartung. In der Studie (ebd.) wurden die Verständlichkeit, Findung des Inhalts und Dauer der Aufgabenlösung in mit SE-kontrollierten vs. unkontrollierten Szenarien mithilfe von Englischmuttersprachlern vs. -nichtmuttersprachlern untersucht und verglichen.

Obwohl die Usability-Technik eine informative Evaluationsmethode ist, die ermöglicht, die MÜ aus einer anderen Perspektive zu betrachten, müssen Aspekte der ökologischen Validität bei der Durchführung solcher Usability-Studien berücksichtigt werden. Unter diese Aspekte fallen die Verwendung von authentischen Texten, der Einsatz von realen bzw. potenziellen Endnutzern sowie die Durchführung im Rahmen von möglichst realitätsnahen Testsettings (vgl. King 1996; Spyridakis u. a. 2005; Doherty & O'Brien 2012). Shubert u. a. (1995) untersuchten zwar Originaldokumentationen der Flugzeugwartung, die Probanden aber waren keine realen Nutzer solcher Dokumentationen. Außerdem kritisierten Holmback u. a. (1996) die Verwendung von Verfahrensdokumentationen zur Bewertung der Verständlichkeit. Sie argumentierten, dass „taking a comprehension test is probably not the best way to measure comprehension of a procedure […] since procedures are written to be performed, not quizzed" (ebd.: 176).

In der vorliegenden Studie konnte die Usability-Technik zur Evaluation aus mehreren Gründen nicht angewendet werden. Erstens handelt es sich bei der Studie um eine Untersuchung auf Satzebene, sprich, es wurden keine längeren vollständigen Texte analysiert, zu denen Aufgaben gelöst und Faktoren wie „goal completion, effectiveness, efficiency, and user satisfaction" im Sinne der Usability-Definition gemessen werden konnten. Zweitens, bestand der Untersuchungsgegenstand einerseits aus authentischen technischen Dokumentationen; andererseits waren diese eine breite Mischung aus Benutzerhandbüchern und Bedienungsanleitungen von Software, Maschinen, Haushaltsgeräten usw. (siehe §4.5.3.1

[1]), für die es kompliziert wäre, reale bzw. potenzielle Nutzer in realitätsnahen Testsettings zu rekrutieren, um ein akzeptables ökologisches Validitätsniveau realisieren zu können.

3.4.3.4.2 MÜ-Evaluation mithilfe des Eyetracking

Ein weiterer interdisziplinärer Ansatz der MÜ-Evaluation findet sich in der Verwendung von Eye-Trackern. Das Eye-Tracking (Blickaufzeichnung oder Blickregistrierung) ist eine Beobachtungsmethode, in der „mittels einer Augenkamera die Blickbewegungen eines Probanden aufgezeichnet werden können" (Stoessel 2002: 80). Das Eye-Tracking (ET) wurde vom Arbeitsbereich Usability-Engineering der Universität des Saarlandes (ABEUS 2006) folgendermaßen definiert „Verfahren, die den Blickverlauf einer Person beim Betrachten eines Bildes registrieren und festhalten". Großflächig wird das ET im Bereich der Usability verwendet. Zudem profitieren viele Studien in anderen Bereichen, wie der Kognitionswissenschaft, Psychologie, Marketing, Mensch-Computer-Interaktion (HCI) im Allgemeinen, und auch in verwandten Bereichen wie der Linguistik, Psycholinguistik und Translationswissenschaft, von diesem nützlichen Beobachtungstool. Die Aufnahme von Informationen zerfällt in „Fixationen" (die Stellen, die der Proband mit den Augen fixiert) und „Sakkaden" (die Augensprünge zwischen den Fixationen).[19] Eine grundlegende Idee im ET ist, dass „das Auge die Einheit fixiert, die gerade vom Gehirn verarbeitet wird" (Funke 2006). Man spricht in der Kognitionswissenschaft von der Eye-Mind-Hypothese, wonach „die Fixation und das Verarbeiten der Informationen im Gehirn eng miteinander verknüpft sind" (ebd.).

Vor diesem Hintergrund werden Eye-Tracking-Messungen, wie die Anzahl der Fixationen, die durchschnittliche Fixationsdauer und die prozentuale Veränderung der Pupillendilatation, als Indikatoren für die kognitive Belastung betrachtet (Duchowski 2007). Andererseits argumentieren Van Gog u. a. (2009), dass Augenmessungen – wie z. B. die Fixationsdauer – andere Aspekte der kognitiven Verarbeitung widerspiegeln können, und begründen dies wie folgt: „[E]xperiencing an overall lower load […] might allow one to allocate more cognitive capacity to processing information in the learning task, thereby leading to higher fixation duration" (ebd.: 328). Nach Van Gog et al. (ebd.) liefert das ET primär

[19] Technisch gesehen ist die Fixation (Jacob & Karn 2003) eine relativ stabile Auge-in-Kopf-Position innerhalb eines gewissen Schwellenwerts von Dispersion (in der Regel ca. 2°) mit einer Mindestdauer (typischerweise 100–200 ms) und einer Geschwindigkeit unterhalb eines bestimmten Schwellenwerts (typischerweise 15–100 Grad pro Sekunde).

Informationen zur visuellen Aufmerksamkeit; „relating visual attention to cognitive processes should be done with great caution, because there is not always a one-to-one relationship (e.g., exogenous shifts in attention may occur unintentional)“ (ebd.: 329). Daher wird empfohlen das ET mit Think-Aloud-Protokollen[20] zu kombinieren, um mehr Einblick in die zugrundeliegenden kognitiven Prozesse zu erhalten (ebd.).

Das ET fand breite Anwendung in der Lesbarkeitsforschung, z. B. bei der Untersuchung der Wort- und Textkomplexität, anhand der Fixationsdauer, Veränderung der Pupillendilatation bzw. Gaze-Time[21] (vgl. Kliegl u. a. 2004; Jensen 2009). Ein weiterer Anwendungsbereich des ET liegt in der Prozessanalyseforschung – sowohl von Humanübersetzungen als auch beim Post-Editing maschineller Übersetzungen – in der das Verhalten der Übersetzer bzw. der Post-Editoren genauer untersucht wird (vgl. Carl u. a. 2015; Hansen-Schirra u. a. 2017).

Im Bereich der MÜ-Evaluation stellten Doherty & O'Brien (2009) die Frage in einer Studie mit dem Titel „Can MT output be evaluated through eye tracking?“ und gingen bei der Antwort folgendermaßen vor. Sie ließen MÜ-Sätze von Humanübersetzern bewerten. In der Bewertung wurden 55 Sätze als gut und 25 Sätze als schlecht übersetzt beurteilt. Daraufhin hatten zehn Muttersprachler in einem ET-Experiment die Aufgabe, die Sätze zu lesen und zu verstehen. Die Ergebnisse zeigten, dass die Gaze-Time bei den schlechten MÜ-Sätzen signifikant höher ausfiel, wobei es bei der Anzahl der Fixationen keinen signifikanten Unterschied gab. (ebd.)

In ihrer Studie „How do humans evaluate Machine Translation?“ tauchten Guzmán u. a. (2015) in den MÜ-Evaluationsprozess ein, indem sie das Verhalten der Bewerter näher analysierten. Sie verglichen die Evaluationsdauer sowie die Konsistenz in der Evaluation von 300 MÜ-Sätzen in zwei Gruppen mit monolingualen und bilingualen Probanden mit unterschiedlichen Verfügbarkeitsszenarien des Ausgangs- und Zieltexts. Die Evaluation fand in Form eins Scorings auf einem 0-100-Schieber statt. Mithilfe des Eye-Trackers wurde die Gaze-Time bei den verschiedenen Szenarien gemessen. Die Ergebnisse zeigen, dass monolinguale Probanden langsamer aber konsistenter bei der Evaluation als bilinguale sind, insb. wenn nur der Zieltext verfügbar ist. Wenn sowohl der Ausgangs- als auch der Zieltext verfügbar sind, nimmt die Evaluation im Allgemeinen mehr Zeit in Anspruch und die Konsistenz bei den monolingualen Probanden sinkt. Entspre-

[20] Beim Think-Aloud denken die Probanden laut während sie die Aufgaben ausführen. Es gibt zwei Arten der Think-Aloud-Protokolle: simultanes Think-Aloud, das die Schritte beleuchtet, die zu einer Entscheidung führen; und retrospektives Think-Aloud, das mehr Details zur Entscheidung liefert. (Hannu & Pallab 2000)

[21] Die Gaze-Time (auch „observation length“) ist die Dauer der Augenbewegungen innerhalb eines Interessenbereichs (Area of Interest „AOI“) (Doherty 2012: 47, 113).

chend empfehlen Guzmán et al. (ebd.) die Durchführung von MÜ-Evaluationen durch monolinguale Probanden und ihnen nur den Zieltext zur Verfügung zu stellen, um eine höhere Konsistenz und Kosteneffizienz zu erreichen. (ebd.)

In einer weiteren Studie analysierte O'Brien (2010a) die Lesbarkeit mithilfe des ET. Sie untersuchte in vier Texten, deren Lesbarkeit im Vorfeld mit Flesch Reading Ease-Formel[22] gemessen wurde, die Fixationsanzahl und -dauer. Die Hypothese hierbei ist, dass Texte, die nach der Lesbarkeitsformel einfach zu lesen sind, mit weniger Fixationen und kürzerer Fixationsdauer gelesen werden. Diese Hypothese konnte nicht für alle Texte belegt werden. Zudem analysierte O'Brien (ebd.) den Einfluss des Einsatzes von KS-Regeln auf die Lesbarkeit eines einfachen Texts und eines komplexen Texts. Beide Texte waren kurz; sie bestanden durchschnittlich aus 192 Wörtern. O'Brien (ebd.) verglich bei jedem Text zwei Versionen, kontrolliert und unkontrolliert, mit der Hypothese, dass die kontrollierten Texte im Vergleich zu den unkontrollierten Texten einfacher zu lesen sind. Nur bei dem komplexen Text konnte eine verbesserte Lesbarkeit im Falle der kontrollierten Version mithilfe des ET durch eine signifikant geringere Fixationsanzahl und einen marginalen Rückgang der Fixationsdauer beobachtet werden. Für den einfachen Text konnte anhand der gleichen ET-Messungen keine Verbesserung der Lesbarkeit durch den Einsatz der KS-Regeln belegt werden.

In seiner Dissertation untersuchte Doherty (2012) den Einfluss der KS auf die Lesbarkeit und Verständlichkeit der MÜ u. a. mithilfe des ET. Das analysierte Korpus wies insgesamt 33 Verstöße auf (ebd.: 103). Doherty (ebd.: 190) analysierte verschiedene Eye-Tracking-Messungen in seiner Pilotstudie und Hauptstudie und stellte Inkonsistenzen in den Korrelationen zwischen den ET-Messungen in den beiden Studien fest, z. B. korrelierten die Gaze-Time und Fixationsanzahl gut in der Pilotstudie, wobei ihre Korrelation in der Hauptstudie nicht signifikant war. Die ET-Ergebnisse zeigen, dass die Fixationsanzahl, die Fixationsdauer, der Regressionsabstand und die Regressionsanzahl[23] nach dem Einsatz der KS sanken (ebd.: 222). Doherty interpretiert diesen Rückgang in den ET-Messungen mit weniger kognitivem Aufwand beim Lesen im Falle des KS-Einsatzes. Gleichzeitig waren nicht alle Differenzen in den ET-Messungen vor vs. nach dem KS-Einsatz signifikant.[24] (ebd.)

[22] Flesch Reading Ease ist eine der am häufigsten verwendeten und getesteten Lesbarkeitsformeln (DuBay 2004).

[23] Eine Regression ist „any eye movement that begins at the right-most point the reader has fixated and leaves the currently fixated region to the left" (Pickering & Traxler 1998: 945). Doherty (2012: 114) berechnete die Regressionsanzahl auf Satzbasis, wobei der Regressionsabstand „the distance travelled for each regression in units of words" darstellt (ebd.).

[24] Die Ergebnisse der umfangreichen Untersuchungen von Doherty (2012) sind unter §3.5 dargestellt.

In der vorliegenden Studie wurde das ET nicht eingesetzt. Anders als die oben dargestellten KS-Studien beschäftigt sich die Studie mit dem Einfluss *einzelner* KS-Regeln und nicht der KS *im Allgemeinen*. Wie unter §4.5.3.1 [9] erklärt, mussten die MÜ-Sätze außerhalb der KS-Stelle korrigiert bzw. vereinheitlicht werden, um die Unterschiede innerhalb der KS-Stelle vor und nach dem KS-Einsatz vergleichen zu können. Aus Gründen der ökologischen Validität wäre es ideal, die MÜ-Sätze nicht zu bearbeiten bzw. zu vereinheitlichen. Jedoch kann solche Bearbeitung auch beim Einsatz vom ET nicht vermieden werden. Tabelle 3.1 veranschaulicht diese Vorgehensweise und ihre Begründung:

Tabelle 3.1: Beispiel 1

vor KS	Ist nur ein Gerät angeschlossen, so ist die Funktion **Punkt-zu-Punkt-Verbindung** zu wählen.
nach KS	Ist nur ein Gerät angeschlossen, so ist die Funktion **“Punkt-zu-Punkt-Verbindung”** zu wählen.
vor KS FehAnno	If only one device is connected, then the function **point-to-point connection** to choose .
nach KS FehAnno	If only one device is connected, then the function "**Point-to-point connection**" VERB .
vor KS HuEv	If only one device is connected, the function **point-to-point connection** must be selected.
nach KS HuEv	If only one device is connected, the function "**Point-to-point connection**" must be selected.

KS-Stelle ist **fett** dargestellt; Fehler innerhalb der KS-Stelle sind unterstrichen; Fehler außerhalb der KS-Stelle sind farblich markiert.

In diesem Beispiel geht es um die KS-Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“. Dementsprechend ist die KS-Stelle ‚Punkt-zu-Punkt-Verbindung‘ im Ausgangssatz bzw. ‚Point-to-point connection‘ im Zielsatz. Wie die Fehlerannotation zeigt, beinhaltete die MÜ vor KS einen Fehler innerhalb der KS-Stelle (das Kleinschreiben in ‚point‘) und einen Fehler außerhalb der KS-Stelle (ein falsches Verb in ‚to choose‘). Ebenfalls im Nach-KS-Szenario gab es einen Fehler außerhalb der KS-Stelle (ein fehlendes Verb). Das Beibehalten der Fehler außerhalb der KS-Stelle würde die Bewerter ablenken. Beim ET würden sie eher die schwerwiegenden bzw. auffälligen Fehler fixieren, wobei

diese Fehler für die KS-Regel irrelevant sein können. Daher würde ein Einsatz des ET beim Studiendesign nicht dazu beitragen, auf die Korrektur und Vereinheitlichung der MÜ-Sätze außerhalb der KS-Stelle zu verzichten und somit die ökologische Validität zu maximieren.

Im Vergleich zu den oben dargestellten ET-Studien wurden in dieser Studie deutlich mehr Sätze (1.100 Sätze) untersucht. Zudem wurden drei Qualitätsparameter (Genauigkeit, Klarheit und Stil) bei der Evaluation analysiert. Gleichzeitig zeigen die vorherigen KS-Studien, dass die ET-Messungen in mehreren Fällen keine signifikanten Ergebnisse liefern. Daher wäre eine Kombinierung des ET mit Think-Aloud-Protokollen – wie z. B. von Van Gog u. a. (2009) empfohlen – erforderlich, um herauszufinden, ob und wie der Bewerter die KS-Stelle bemängelt. Ein solches Experimentdesign hätte den Zeit- und Kostenaufwand der Evaluation erhöht. Anders als von Guzmán u. a. (2015) empfohlen konnte die Evaluation von monolingualen Bewertern mit der bloßen Anzeige der Zielsätze zur Kosteneinsparung und Konsistenzerhöhung beim ET nicht durchgeführt werden. Die Evaluation der genannten MÜ-Qualitätsparameter setzte translatorische Kompetenzen voraus. Daher musste die Evaluation mithilfe von qualifizierten Übersetzern durchgeführt werden. Zudem war die Anzeige der Ausgangssätze zur Bewertung der Genauigkeit erforderlich.[25] Diese Settings hatten keinen negativen Einfluss auf die Konsistenz, denn das Interrater-Agreement lag durchschnittlich bei 0,840 (siehe §5.2.3, Tabelle 5.2). Ein ET-Experiment, in dem die einzelnen Qualitätsparameter in gesonderten Screens bewertet würden, hätte den Zeit- und Kostenaufwand erheblich erhöht. Aus diesen Gründen wurde die Evaluation in Form eins Scorings auf Likert-Skalen kombiniert mit Checkbox und Post-Editing durchgeführt (siehe Abbildung 4.6). Anhand der Checkboxen konnten die vergebenen Scores nach konkreten Qualitätskriterien begründet werden. Zudem lieferte das im Rahmen der Humanevaluation durchgeführte Post-Editing wertvolle Inputs bei der qualitativen Analyse der Ergebnisse. Die Evaluationskosten ließen sich eindämmen und nicht zuletzt musste die Evaluation nicht im Labor durchgeführt werden, sondern die Bewerter genossen Flexibilität bei der Auswahl der Zeit und des Orts der Evaluationsdurchführung, was für die ökologische Validität förderlich war.[26]

[25] Für mehr dazu siehe „Darstellung der Ausgangssätze“ unter §4.5.5.2.

[26] Für mehr zum Evaluationsdesign und -ablauf siehe „Ablauf der Evaluation“ unter §4.5.5.2.

3.5 MÜ-Evaluation im Kontext der Kontrollierten Sprache

In diesem Abschnitt werden Evaluationsstudien, die den Effekt der Kontrollierten Sprache (KS) auf den MÜ-Output bei verschiedenen MÜ-Ansätzen mit diversen Evaluationstechniken untersucht haben, dargestellt. Zudem werden die niedrige Frequenz von KS-Untersuchungen auf Regelebene und der Sinn solcher Untersuchungen thematisiert. Zum Schluss werden die Herausforderungen einer KS-Untersuchung zusammen mit einer Erläuterung des Umgangs mit ihnen im Rahmen dieser Studie diskutiert.

3.5.1 Studien zur MÜ-Evaluation im Kontext der KS

Wie bereits (unter §2.2.2) dargestellt, unterscheidet man bei den Kontrollierten Sprachen zwischen einer humanorientierten KS (HOCL) und einer maschinenorientierten KS (MOCL) (vgl. Huijsen 1998). Eine HOCL hat das Ziel, die Lesbarkeit und Verständlichkeit des Texts, in der Regel von englischen technischen Dokumentationen für den internationalen Markt, d. h. für Nichtmuttersprachler, zu verbessern und gleichzeitig die Übersetzungskosten zu minimieren. Bei einer MOCL hingegen liegt der Schwerpunkt auf der Verbesserung der Übersetzbarkeit durch ein Pre-Editing des Ausgangstexts. Bei beiden Gruppen basiert die dahinterstehende Idee darauf, dass die KS die Ambiguität reduziert bzw. auflöst und die Komplexität der Satzstruktur einschränkt (Hutchins & Somers 1992: 4; O'Brien 2006). Bei den MOCL profitiert die MÜ insbesondere von den Regeln zur Verkürzung der Satzlänge und Eliminierung der sog. „negative translatability indicators" (NTIs). NTIs sind „features that are known to be problematic for MT" (Underwood & Jongejan 2001; O'Brien 2006: 1), wie z. B. das Gerundium im Englischen, missverständliche Genitivkonstruktionen und Funktionsverbgefüge im Deutschen und mehrdeutige pronominale Bezüge in beiden Sprachen.

Die Regeln der MOCL sind meist umfangreicher und strenger als die der HOCL, da erstere primär darauf abzielt, die Ambiguität zu reduzieren (Mitamura 1999). In einer Studie mit dem Titel „Two in one – Can it work?" bildete Reuther (2003) zwei KS-Regelsätze: den ersten Regelsatz zur Verbesserung der Lesbarkeit und Verständlichkeit und den zweiten zur Verbesserung der Übersetzbarkeit. Sie fand heraus, dass die beiden Regelsätze sich nicht allzu sehr unterschieden; die Regeln der Lesbarkeit und Verständlichkeit waren eine Teilmenge der Regeln der Übersetzbarkeit (ebd.). Die Ergebnisse von Reuther (ebd.) korrelieren teilweise mit den Schlussfolgerungen von Bernth & Gdaniec (2001). In ihrer Arbeit prägten Bernth und Gdaniec (ebd.) den Begriff „MTranslatability". MTranslatability

steht für die Übersetzbarkeit von Texten, die mithilfe von MÜ-Systemen übersetzt werden (ebd.: 175). Ziel der Arbeit war, Merkmale zu identifizieren, die die maschinelle Übersetzbarkeit beeinträchtigen, und entsprechend Regeln zu ihrer Verbesserung vorzuschlagen. Sie definierten 26 Regeln für die englische Sprache und gruppierten sie unter: grammatische Strukturen, ambige Strukturen, stilistische Aspekte, orthografische Aspekte und Dateieigenschaften.

Unter den ersten KS, die in Verbindung mit einem bestimmten MÜ-System im Bereich der technischen Dokumentation verwendet wurden und zur wesentlichen Reduzierung der Übersetzungskosten sowie Steigerung der Produktionseffizienz der Dokumentation führten, waren Xerox Multinational Customized English in Verbindung mit Systran (damals ein RBMÜ-System) (Elliston 1979) und Caterpillar Technical English in Verbindung mit dem RBMÜ-System KANT (Nyberg & Mitamura 1996; Kamprath u. a. 1998). Untersuchungen von KS in Zusammenhang mit der MÜ werden in der Regel von Unternehmen durchgeführt, um die Effektivität der verwendeten Regeln zu bewerten und sie entsprechend zu optimieren. Aus Vertraulichkeitsgründen werden jedoch die Ergebnisse nicht genauer angegeben (vgl. Bernth & Gdaniec 2001: 207) oder die implementierten Regeln nicht vollständig bzw. nur anhand von Beispielen veröffentlicht. Dennoch soll im Folgenden ein Überblick über die wesentlichen Studien präsentiert werden.

Der positive Einfluss der KS auf die MÜ ist nicht nur bei den ersten KS zu beobachten. Mehrere theoretische und empirische Studien geben an, dass die Durchführung von Pre-Editing in Form von KS-Regeln zur Verbesserung der Qualität des MÜ-Outputs führt. (vgl. Nyberg & Mitamura 1996; Bernth 1999; Bernth & Gdaniec 2001: 208; Drugan 2013: 98; Drewer & Ziegler 2014: 196; Wittkowsky 2017: 92) Eine der häufig zitierten Studien ist die Untersuchung von Nyberg & Mitamura (1996), in der sie zunächst die Architektur des verwendeten MÜ-Systems „KANT" (ein knowledge-basiertes MÜ-System) sowie die Ambiguitäten im Ausgangstext in der technischen Domäne analysiert haben. Anschließend untersuchten sie die zusammen mit KANT angewendeten KS-Regeln und zeigten, dass die Genauigkeit des MÜ-Outputs stark vom Steuerungsniveau des Ausgangstexts abhängt. Bernth (1999) untersuchte den Einfluss von KS-Regeln auf den MÜ-Output vom MÜ-System LMT[27] in der technischen Domäne mithilfe einer Humanevaluation. Die Ergebnisse zeigten unter Berücksichtigung der KS-Regeln eine deutliche Verbesserung der Übersetzungsqualität. In einer weiteren bekannten Studie von Bernth & Gdaniec (2001: 208) wurde eine Reihe von

[27] LMT (Logical Based Machine Translation) ist ein experimentelles Englisch > Deutsch MÜ-System, das im Rahmen der Logikprogrammierung entwickelt wurde (McCord 1989).

KS-Regeln bei englischen Fragen und Antworten (Q&A) im Bereich der Pflanzenpflegeanweisungen umgesetzt. Anschließend wurden sie vor und nach der Umsetzung der KS-Regeln von vier MÜ-Systemen ins Französische, Deutsche und Spanische übersetzt. Auf Basis einer Evaluation von drei Muttersprachlern in jeder Zielsprache wurde der MÜ-Output auf Dokument- und Satzebene in Hinsicht auf das Kriterium Verständlichkeit bewertet. Sowohl auf Dokument- als auch auf Satzebene stieg die Qualität der MÜ nach dem Einsatz der KS-Regeln.

Der Zusammenhang zwischen dem Pre-Editing und dem Post-Editing (PE) wurde von einigen Studien untersucht. Das Pre-Editing zielt nach O'Brien (2010b) auf „modifying the input text before automatic translation to facilitate machine processing" ab. Pre-Editing-Techniken können vereinfacht in Form von Style-Guides oder als kontrollierte Terminologie und KS-Regeln umgesetzt werden (vgl. ebd.). Diese wiederum ermöglichen den MÜ-Systemen eine bessere Verarbeitung des Ausgangstexts, daher wird das Pre-Editing zur Verbesserung des MÜ-Outputs empfohlen (vgl. Drugan 2013: 98). Auf der anderen Seite wird die MÜ mit dem Ziel posteditiert (siehe §3.4.3.3), den MÜ-Output zu korrigieren bzw. zu verbessern (TAUS Report 2010). Somit haben beide Techniken, Pre-Editing und Post-Editing, die Optimierung des MU-Outputs als gemeinsames Ziel. Schafft das Pre-Editing des Ausgangstexts eine einfache Satzstruktur und eindeutige Inhalte, so kann das PE mehrerer Zieltexte effizienter ausfallen. Vor diesem Hintergrund befassten sich einige Studien mit der Fragestellung, ob ein Pre-Editing in Form des Einsatzes der KS zur Beschleunigung oder Erhöhung der Effektivität des PE führen kann. Im Folgenden werden Studien in diesem Bereich dargestellt.

Aikawa u. a. (2007) untersuchten die Relationen zwischen der KS, der MÜ und des PE in der technischen Domäne. Auf Basis der Ergebnisse von früheren Studien, die zeigen, dass die KS die Qualität der MÜ verbessert, war das Ziel von Aikawa et al. (ebd.) herauszufinden, ob der Einsatz der KS zur höheren Produktivität oder zum geringeren PE-Aufwand führt, wodurch eine Dreierrelation zwischen der KS, MÜ und PE geschaffen werden kann. Zwei Datensätze inkl. KS-konformer Sätze und KS-nicht-konformer Sätze wurden mithilfe eines SMÜ-Systems (MSR-MT von Microsoft) übersetzt. Die MÜ-Version sowie die post-editierte Version der beiden Datensätze wurden in drei Analysen verglichen: einer Humanevaluation (auf einer 4er-Skala), einer automatischen Evaluation (auf Basis von BLEU-Scores) und einem Vergleich der zeichenbasierten Edit-Distances. Die Ergebnisse zeigten, dass der Einsatz der KS sowohl die PE-Produktivität als auch die MÜ-Qualität verbessert. Eine weitere Studie (Temnikova 2010) verglich den kognitiven PE-Aufwand bei der Identifizierung und Korrektur von verschiedenen MÜ-Fehlertypen in zwei Fällen: MÜ von kontrollierten Texten und MÜ von nicht-kontrollierten Texten. Temnikova (ebd.) kam

ebenfalls zu positiven Ergebnissen im Falle der kontrollierten Texte, nämlich dass deren MÜ Folgendes beinhaltet: (1) weniger Fehler, die kognitiv schwer zu identifizieren und beheben sind und (2) mehr Fehler, die kognitiv leicht zu erkennen und korrigieren sind. Thicke (2011) untersuchte den Einfluss von KS-Regeln des Global English auf den MÜ-Output in Bezug auf die PE-Produktivität i. S. v. PE-Zeitaufwand. Die Studie wurde in der technischen Domäne und mit dem hybriden System Systran durchgeführt (ebd.). Die PE-Zeit wurde in vier Szenarien verglichen: untrainiertes MÜ-System vor vs. nach dem Einsatz der KS-Regeln und trainiertes MÜ-System ebenfalls vor und nach dem KS-Einsatz. Die Ergebnisse zeigen, dass (1) ein PE bei untrainiertem System mit einem Datensatz vor dem KS-Einsatz doppelt so produktiv (d. h. schnell) wie eine Humanübersetzung ist; (2) ein PE bei trainiertem System aufgrund des Vorhandenseins der Terminologie vor dem KS-Einsatz dreifach so produktiv wie eine Humanübersetzung ist; (3) ein PE bei trainiertem System nach dem KS-Einsatz vierfach so produktiv wie eine Humanübersetzung ist, da die KS-Regeln zur Behebung von grammatischen Fehlern beitrugen. (ebd.)

Während die meisten MÜ-Evaluationsstudien im Bereich der KS sich auf die technische Domäne fokussieren, wurden im Rahmen des Forschungsprojekts ACCEPT (Automated Community Content Editing PorTal) Inhalte technischer Foren (user-generated content (UGC)) untersucht (Lehmann u. a. 2012; Gerlach u. a. 2013). Ziel des ACCEPT-Projekts war die Verbesserung des SMÜ-Outputs dieser Textsorte durch minimales Pre-Editing, SMÜ-Verbesserungsmethoden sowie PE-Strategien (Gerlach u. a. 2013). Anders als die technischen Dokumente zeichnen sich Texte technischer Foren dadurch aus, dass sie nicht standardisiert sind, einen informellen bzw. gesprochenen Stil haben und oft alternative Schreibweisen, Akronyme und umgangssprachliche Wörter beinhalten. Ein Systemtraining in dem Fall ist mangels einer genügenden Anzahl an übersetzten Forentexten, die als Trainingsdaten eingesetzt werden könnten, nicht möglich (Lehmann u. a. 2012). Lehmann et al. (ebd.) trainierten das System mit technischen Dokumentationen; das Ziel des PE war es, den Ausgangstext vergleichbar zu den technischen Dokumentationen zu standardisieren. Die Studie verwendete den CL-Checker Acrolinx einerseits zur Kontrolle des Ausgangstexts, andererseits zum PE des Zieltexts. In der Evaluation verglich die Studie anhand von AEMs (BLEU, GTM und TER) die MÜ-Qualität vor und nach dem Einsatz jeder einzelnen Regel und zeigte eine deutliche Verbesserung der MÜ-Qualität nach dem Einsatz der Pre-Editing-Regeln. (ebd.) Gerlach u. a. (2013) untersuchten, ob Pre-Editing-Regeln, die MÜ-Qualität eines SMÜ-Systems verbessern und ob sie sich positiv auf die PE-Produktivität auswirken. Auf Grundlage des PE-Zeitaufwands zeigte die Studie, dass Pre-Editing-Regeln, die sich positiv auf die MÜ-Qualität auswirken, die

PE-Zeit signifikant beschleunigen bzw. fast um die Hälfte reduzieren. Auf Basis dieses Ergebnisses argumentieren Gerlach et al. (ebd.), dass die Kombinierung der Pre-Editing- und Post-Editing-Techniken eine vielversprechende Methode ist, da ein einfacher einsprachiger Pre-Editing-Aufwand den mühsamen zweisprachigen PE-Aufwand effektiv reduziert.

Nicht immer zeigen die KS-Studien eine positive Wirkung auf die MÜ. Es existieren Studien, die auf Basis unterschiedlicher Evaluationsmethoden verschiedene Auswirkungen der KS bzw. keine signifikante Wirkung zeigten. O'Brien (2010a) untersuchte den Einfluss der KS auf die Akzeptanz und Lesbarkeit der MÜ in drei Zielsprachen (Französisch, Spanisch und Chinesisch) bei einem einfachen und einem komplexen Text. Anhand eins Scorings zeigen die Ergebnisse nur bei dem komplexen Text einen kleinen positiven Effekt der KS auf die Lesbarkeit und Akzeptanz der MÜ. Eine Fehleranalyse, in der die kontrollierten und unkontrollierten Versionen der Texte verglichen wurden, konnte weder im einfachen noch im komplexen Text eine Verbesserung des MÜ-Outputs nach dem Einsatz der KS-Regeln belegen. Doherty (2012) analysierte den Einfluss der KS auf die Lesbarkeit und Verständlichkeit des MÜ-Outputs. Die Studie wurde mit einem hybriden MÜ-System (MaTrEx)[28] für die Übersetzung aus dem Englischen ins Französische durchgeführt. Der englische technische Text wurde mithilfe des CL-Checkers Acrocheck kontrolliert. Die MÜ der kontrollierten und unkontrollierten Texte wurde nach einem umfangreichen Mixed-Methods-Ansatz mit den folgenden Ergebnissen verglichen (ebd. 216f.): Die unkontrollierten Varianten beinhalteten mehr Fehler mit leicht unterschiedlichen Fehlertypen. Die Lesbarkeitscores (Flesch und LIX) der kontrollierten Variante waren zwar höher, die Differenz war aber nicht signifikant. Die Differenz in den Scores der AEMs (GTM, BLEU und TER) war ebenfalls nicht signifikant. Bezüglich der ET-Messungen gab es bei der kontrollierten Variante signifikant weniger Fixationen und eine kürzere Fixationsdauer, während alle weiteren ET-Messungen (Gaze-Time, Pupillendilatation und Regressionen) keine signifikanten Unterschiede aufwiesen. Bei der Humanevaluation waren alle Scores (Lesbarkeit, Verständlichkeit und Recall) im Fall der kontrollierten Variante signifikant höher.

3.5.2 Notwendigkeit der KS-Untersuchungen auf Regelebene

Wie die Ergebnisse der vorherigen Studien zeigen, ist die MÜ des kontrollierten Texts nicht immer eindeutig besser als die des unkontrollierten. Eine Untersuchung auf Regelebene könnte die Wirkung der einzelnen Regeln aufdecken.

[28] MaTrEx ist ein hybrides MÜ-System aus SMÜ- und beispielbasierten MÜ-Komponenten (Groves 2007).

In der Praxis sind KS-Untersuchungen auf Regelebene zweifellos aus mehreren Gründen hilfreich. *Im Folgenden wird das Rationale solcher Untersuchungen erläutert*:

(1) Nicht immer wirken sich die KS-Regeln positiv aus. Die KS-Regeln können auch Nebenwirkungen haben, z. B. in Form von beeinträchtigter Textakzeptanz (Roturier 2006). Nyberg u. a. (2003: 257) betonen, dass „some writing rules may even do more harm than good". Es wäre hilfreich, zu erfahren, welche Regeln am effektivsten sind, welche als optional eingestuft werden können und welche einen potentiell negativen Effekt (z. B. stilistisch) hätten und entsprechend in bestimmten Fällen eingesetzt bzw. vermieden werden sollten. O'Brien (2003) verglich acht KS-Regelsätze mit dem Ziel einen Kernregelsatz festzulegen, mit dem die Unternehmen beim Einsatz der KS das Rad nicht neu erfinden müssen. In den acht KS-Regelsätzen gab es nur eine einzige gemeinsame Regel, nämlich „Sätze kurz zu halten (max. 23 Wörter)". Sie begründete diese fehlende Gemeinsamkeit mit der Individualität der Regelsätze, die je nach Ziel des KS-Einsatzes, eingesetztem MÜ-System, Übersetzungsrichtung, Einfluss der Corporate-Schreibregeln bzw. Autoren und Subjektivitätsgrad bei der Definition des Regelsatzes variieren (ebd.). Dementsprechend kann eine systematische Untersuchung der einzelnen Regeln die Nützlichkeit und Notwendigkeit jeder Regel in bestimmten Settings aufdecken, was wiederum die Unternehmen dabei unterstützt, ihre Individualität zu fördern.

(2) In vielen Unternehmen sind die Fachabteilung oder Produktmitarbeiter für die Dokumentation verantwortlich. Diese Mitarbeiter sind zwar häufig Experten im jeweiligen Service oder Produkt, jedoch verfügen sie oft über unzulängliches linguistisches Wissen bzw. sind nicht mit linguistischen Jargons vertraut, um alle Regeln verstehen und umsetzen zu können, was gelegentlich zu Verwirrung führt (Van der Eijk u. a. 1996; Aranberri & Roturier 2009). Sie wollen den minimalen Regelsatz lernen, der sprachübergreifend den größten Einfluss auf die MÜ-Qualität hat (Aikawa u. a. 2007). Daher empfiehlt sich eine Reduktion der Regeln auf das Wesentliche.

(3) KS-Usability und Autorenproduktivität sind von besonderer Bedeutung beim Einsatz der KS. Eine KS, die zur Optimierung der Übersetzung Satzstrukturen und Vokabeln sehr restriktiv einschränkt, kann den Autoren Usability- und Produktivitätsprobleme verursachen (Mitamura 1999). „The

rules imposed by a CL can reduce or force expression", was den Schreibprozess übermäßig komplex gestaltet und zur Ablehnung durch die Autoren führt (Doherty 2012: 31). Findet das Unternehmen heraus, welche Regeln für seine Zwecke unerlässlich sind und welche optional sein können, ermöglicht dies ihm, den Autoren mehr Flexibilität bei der Umsetzung der KS anzubieten. Gleichzeitig ist die Umsetzung von wenigen Regeln wünschenswert, um den Schreibprozess nicht übermäßig zu beeinflussen (O'Brien & Roturier 2007).

(4) Der Einsatz eines großen KS-Regelsatzes kann aus Zeit- und Ressourcengründen schwierig ausfallen, selbst wenn ein KS-Checker verwendet wird (Govyaerts 1996; O'Brien & Roturier 2007). Die Implementierung der effektivsten Regeln ist eine vorteilhafte Alternative, die voraussetzt, dass die Wirksamkeit dieser Regel empirisch geprüft wurde (O'Brien & Roturier 2007).

(5) Dadurch, dass die MÜ-Qualität je nach Sprachenpaar, Übersetzungsrichtung, Domäne und MÜ-System variiert, ist zu erwarten, dass die Auswirkung der einzelnen Regeln zusammen mit diesen Variablen ebenfalls variiert. Dementsprechend kann eine Untersuchung der einzelnen Regeln mehrere Aspekte aufdecken, nämlich ob bestimmte Regeln sprachebergreifend/sprachspezifisch, systemübergreifend/-spezifisch bzw. domänenspezifisch/allgemein sind. Insbesondere nach der Entwicklung des neuesten MÜ-Ansatzes der NMÜ, der mithilfe der künstlichen Intelligenz versucht, menschliches Denken nachzuahmen, ist es an der Zeit, den KS-Einsatz wieder aufzugreifen, um herauszufinden, ob und welche Regeln für die maschinelle Übersetzbarkeit dieses vielversprechenden Ansatzes erforderlich sind.

Trotz der Notwendigkeit der KS-Untersuchung auf Regelebene widmete sich nur eine begrenzte Anzahl von Evaluationsstudien diesem Forschungsgebiet. So beschrieben Nyberg u. a. (2003: 257) diesen Forschungsbedarf in Kürze folgendermaßen: „It is unclear what the contribution of each individual writing rule is to the overall effect of the CL". *Im Folgenden werden Studien mit dem Fokus auf der Analyse des Einflusses einzelner KS-Regeln präsentiert.*

Roturier (2006) interessierte sich für die Analyse der Auswirkungen von 54 einzelnen KS-Regeln auf die Verständlichkeit, Nützlichkeit (usefulness) und Akzeptanz der MÜ von webbasierten Inhalten. Der Fokus der Studie lag auf den Sprachenpaaren Englisch > Deutsch und Englisch > Französisch. Der Datensatz

bestand aus 304 Sätzen. Zur Prüfung der KS-Regelkonformität wurde der CL-Checker Acrocheck verwendet. Die Übersetzung wurde mit einem RBMÜ-System (Systran) angefertigt. Die Ergebnisse zeigen, dass eine begrenzte Anzahl von KS-Regeln überlappend einen ähnlichen Einfluss auf die Verständlichkeit des französischen und deutschen MÜ-Outputs auf Segmentebene hat. Das Online-Experiment deckte bestimmte KS-Regeln auf, die die Verständlichkeit der technischen Dokumentation für die deutsche MÜ erheblich verbessern können. Bei der französischen MÜ konnte die Einführung der KS-Regeln zu keiner wesentlichen Verbesserung der Verständlichkeit, der Nützlichkeit oder der Akzeptanz führen.

O'Brien (2006) untersuchte, ob der Einsatz der KS i. S. v. dem Vermeiden von 29 NTIs (negative translatability indicators) den MÜ-Output verbessert. Sie maß den Einfluss auf die MÜ hinsichtlich des PE-Aufwands, indem sie eine „Choice Network Analyse"[29] durchführte und das Tastatur-Monitoring-Tool „Translog" verwendete. Ihr Datensatz bestand aus einem englischen Benutzerhandbuch aus dem IT-Bereich mit 1.777 Wörtern. Der Text wurde mithilfe eines RBMÜ-Systems (IBM WebSphere) ins Deutsche übersetzt und von neun professionellen Übersetzern posteditiert. Die Ergebnisse zeigen, dass der zeitliche und technische Aufwand für die Sätze, die bekannte NTIs enthalten, größer ist. Zudem zeigt die Studie den unterschiedlichen Einfluss der KS-Regeln auf die MÜ i. S. v. unterschiedlich hohem PE-Aufwand bei den einzelnen Regeln. Ferner zeigt die Studie, dass der PE-Aufwand auch für Sätze betrieben werden kann, bei denen NTIs eliminiert wurden.

Ramírez Polo (2012) analysierte den Einfluss von einzelnen KS-Regeln auf die Qualität der MÜ in einer Humanevaluation und automatischen Evaluation (mithilfe der AEMs BLUE und NIST). Sie untersuchte eine korpusbasierte Testsuite aus 149 Sätzen aus der Automobilindustrie. Regelverstöße wurden mithilfe des CL-Checkers MULTILINT[30] identifiziert. Im Fokus der Analyse stand das Sprachenpaar Deutsch > Englisch; die MÜ wurde mit einem hybriden System (Personal Translator) vorgenommen. Ramirez Polo (ebd.) gruppierte die Regeln in vier Gruppen: Grammatik-, Orthografie-, Terminologie- und Stilregeln. Die Ergebnisse zeigen, dass die meisten Regeln, die eine Qualitätsverbesserung bewirkten, Grammatik- und Orthografieregeln waren. Auf der anderen Seite bewirkten die Terminologie- und Stilregeln zwar einen positiven Effekt, der aber niedriger als erwartet ausfiel (ebd.: 274).

[29] Eine „Choice Network Analysis" vergleicht die Wiedergaben einer einzelnen Zeichenfolge von mehreren Übersetzern, um ein Netzwerk von Auswahlmöglichkeiten vorzuschlagen, das theoretisch das kognitive Modell darstellt, das jedem Übersetzer zum Übersetzen dieser Zeichenfolge zur Verfügung steht (Campbell 2000).

[30] MULTILINT ist der Vorgänger von CLAT, siehe §2.6 „CL-Checker".

Siegel (2013) verglich den Einfluss von Pre-Editing-Regeln bei zwei MÜ-Ansätzen, nämlich RBMÜ (Systran) und SMÜ (Google Translate), für die deutsche KS bei einer MÜ ins Englische und Italienische. Bei dem Vergleich des KS-Einflusses auf Systemebene identifizierte Siegel (ebd.) Regeln, die systemübergreifend einen ähnlichen Effekt haben und andere, die systemspezifisch mehr Einfluss zeigen. Auf Sprachenpaarebene konnten keine Unterschiede im Einfluss der Regeln auf den MÜ-Output festgestellt werden, die sich durch die Struktur des Englischen oder Italienischen als Zielsprache erklären lassen.

Um die Evaluation einzelner Regeln effizienter zu gestalten, verglichen Roturier u. a. (2012) den Einfluss der Regeln auf die MÜ-Qualität mithilfe von AEMs (GTM, BLEU und TER). Außerdem führten sie eine Humanevaluation (i. S. v. Scoring als „bessere", „vergleichbare" oder „schlechtere" MÜ) durch. Die Studie wurde für die Sprachenpaare Englisch > Französisch und Englisch > Deutsch mit Inhalten eines technischen Forums (user-generated content (UGC)) und mithilfe eines PBMÜ-Systems (Moses) durchgeführt. Es wurde der CL-Checker Acrolinx verwendet und die Umsetzung der Regeln erfolgte automatisiert. Von den analysierten Regeln gab es für jedes Sprachenpaar Regeln, die einen positiven Einfluss auf die MÜ hatten, andere mit negativem Einfluss und welche, die keinen Einfluss zeigten. Zudem identifizierte die Studie Regeln, die sich überlappend (positiv oder negativ) auf beide Sprachenpaare auswirken. (ebd.)

In einer weiteren Studie, analysierte Møller (2003) fünf Regeln des Simplified English (SE) in einem kleinen Korpus übersetzt aus dem Englischen ins Dänische mithilfe eines RBMÜ-Systems (Comprendium). Ziel ihrer Studie war, zu untersuchen, ob ein Schreibstil, in dem verschiedene Formen grammatischer Metaphern aufgelöst wurden, um den Text für den Leser zugänglicher zu machen, tatsächlich die Eignung der Texte für die MÜ fördert. (ebd.) Allerdings wurde die Studie empirisch nicht untermauert.

3.5.3 Forschungsherausforderungen der KS-Untersuchungen auf Regelebene

Die Bewertung der Auswirkung der KS auf die MÜ ist ein komplexes Thema, bei dem mehrere Variablen berücksichtigt werden müssen. Sie gewinnt an Komplexität, wenn es um die Untersuchung einzelner Regeln geht. Das erklärt die begrenzte Anzahl an Studien in diesem Bereich. Im Folgenden wird auf Details der Forschungsherausforderungen bei diesem Thema eingegangen. Zudem wird erläutert, wie in der vorliegenden Studie mit diesen Herausforderungen umgegangen wurde.

Untersuchungsgegenstand: Die Forscher versuchen in der Regel die Untersuchung mit einem authentischen Datensatz durchzuführen. Eine Herausforderung von authentischen Texten ist, dass ein Satz oft Verstöße gegen mehrere KS-Regeln beinhalten kann. Nicht selten können mehrere Regeln gleichzeitig nicht angewendet werden, denn „resolving a rule violation resulted in further violations" (Doherty 2012: 103). In den Fällen, in denen man mehrere Regeln gleichzeitig anwenden kann, besteht das Problem darin, dass „it was not always straightforward to determine which rule had an effect on the quality of the segment, since many segments were affected by more than one rule" (Ramírez Polo 2012: 274). Beispiel hierfür ist ein Satz, in dem sowohl ein Modalverb als auch Nominalisierung verwendet wird; nach den zwei stilistischen Regeln der tekom e.V.-Leitlinie von 2013 sollten Modalverben und Nominalisierungen vermieden werden. Bei der Analyse eines solchen Satzes kann der Einfluss von nur einer der beiden Regeln nicht untersucht werden. Analysiert man z. B. einen Satz, der zwei Verstöße beinhaltet, wobei sich eine Regel positiv und die andere negativ auswirken würde, erhält man einen gemischten Effekt, dessen genaue Quelle schwer erkennbar ist. Es ist schwer, eine große Anzahl an authentischen Sätzen zu finden, in denen jeweils nur ein Verstoß gegen eine Regel vorliegt. Dementsprechend wurde in der vorliegenden Studie – vergleichbar mit Roturier (2006: 74) – nur die untersuchte Regel in jedem Satz eingesetzt. Ein gleichzeitiger Einsatz mehrerer Regeln würde nicht nur die Realisierung des Untersuchungsziels verhindern, sondern würde auch die Gefahr mit sich bringen, dass „some of the test suite's segments could be contaminated and the internal validity of the study would be undermined" (ebd.: 74).

Die Größe der Datensätze in den präsentierten Studien zeigt, welche Herausforderung es ist, einen ausgewogenen Datensatz im Hinblick auf die Anzahl, Länge und den Schwierigkeitsgrad der analysierten Sätze pro Regel zu haben. Nicht selten müssen die Forscher zwischen der Anzahl der analysierten Regeln und der der analysierten Sätze pro Regel abwägen (ebd.: 55). So analysierte O'Brien (2006) 29 NTIs, wobei die Sätze aus einer Datei (Benutzerhandbuch) stammten. In der Datei waren einige NTIs überrepräsentiert, während andere NTIs gar nicht vorkamen. Für die von Roturier (2006) untersuchten 54 Regeln wurden pro Regel zwei bis 14 Sätze (insgesamt 304 Sätze) analysiert. Des Weiteren berichtet Roturier (ebd.: 55) von zwei weiteren Studien von McCarthy (2015) und Rochford (2005) in diesem Bereich, die kleine Dateisätze untersuchten: McCarthy analysierte 22 Regeln mit einem Datensatz von zwei Sätzen pro KS-Regel, um die Auswirkungen der Regeln auf die MÜ-Qualität von zwei Zielsprachen zu bewerten. Rochford konzentrierte sich auf acht Regeln und verglich dabei den MÜ-Output von drei MÜ-Systemen. Ihr Datensatz umfasste einen bis vier Sätze pro KS-Regel.

Um eine Ausgewogenheit zu ermöglichen, verwenden die Studien in diesem Bereich oft, wie auch der Fall in der vorliegenden Studie, Testsuites. Nach King & Falkedal (1990) sollte eine Testsuite „at least two test inputs for each structure" beinhalten. Manche KS-Regeln lassen sich aber nicht mit nur zwei Sätzen untersuchen, daher bemühen sich die Forscher mehr Sätze zu analysieren (vgl. Roturier 2006: 71). In der vorliegenden Studie wurden z. B. die Regeln „Partizipialkonstruktionen vermeiden" und „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" analysiert. Bei diesen beiden Regeln spielt die Länge der Partizipialkonstruktion bzw. des Oberflächentexts eine Rolle bei der Untersuchung des Einflusses der jeweiligen Regel. Entsprechend muss eine gewisse Anzahl an Sätzen pro Variante im Datensatz vertreten sein, um überhaupt ein signifikantes Ergebnis erreichen zu können. Für eine zuverlässige Auswertung bzw. ein robustes statistisches Ergebnis wurden in der vorliegen Studie 216 Ausgangssätze (24 Sätze * 9 Regeln) bei fünf MÜ-Systemen analysiert. Die Sätze stammen aus zehn Dateien unterschiedlicher Quellen: fünf Bedienungsanleitungen für Haushaltsgeräte; zwei Pflegeanleitungen von zwei Hausausstattungen (Teppich und Küchenmöbel); einer Betriebsanweisung einer Fabrikmaschine; einem Benutzerhandbuch einer Software; einer Gepäckregelung einer Fluggesellschaft. Somit waren alle Quellen mit Ausnahme der Betriebsanweisung der Maschine und des Benutzerhandbuchs der Software für den normalen Endnutzer (ohne bestimmte Fachkenntnisse) bestimmt und inhaltlich wurde keine Komplexität erwartet. Zudem wurden komplexe oder sehr spezifische Fachbegriffe durch bekannte Begriffe ersetzt, um zu vermeiden, dass die Bewerter sich während der Humanevaluation mit solchen Begriffen unnötig intensiv beschäftigen (Genaueres dazu unter §4.5.3.1, Schritt [4]).

Datenaufbereitung: Nachdem der Forscher eine angemessene Anzahl an Sätzen sammelt, folgt ein wichtiger Schritt, die Eliminierung von „noise" (King & Falkedal 1990), i. S. v. eine Reduzierung der linguistischen Schwierigkeiten, die für das getestete Problem irrelevant sind. Dieser Schritt ist nicht einfach, da eine Eliminierung von „noise" nach klaren Kriterien vorgenommen werden muss, um eine Verzerrung der Daten zu verhindern. O'Brien (2006) musste die analysierte Datei bearbeiten, um die Häufigkeit der verschiedenen NTI-Typen auszugleichen (ebd.: 133). Trotz dieser unvermeidlichen Bearbeitung war sie bemüht, dass „the text would still be recognised by the post-editors as a user manual as opposed to a list of artificially constructed sentences" (ebd.: 134). Ebenfalls wurde in der vorliegenden Arbeit angestrebt, die Ausgangssätze nach vordefinierten Kriterien vom „noise" zu befreien (siehe §4.5.3.1 [4]). Zudem wurde die stilistische Akzeptanz in einem Qualitätssicherungsschritt durch zwei unabhängige Bewerter geprüft und bei Bedarf ersetzt (siehe Schritt [6] unter §4.5.3.1).

Umsetzung der KS-Regeln: Sobald der Forscher seinen Datensatz erstellt hat, beginnt der nächste Schritt mit der Umsetzung der KS-Regeln und damit erscheint die nächste Herausforderung, nämlich: „Wie genau werden die Regeln umgesetzt?" Nach Roturier (2006: 74) „[m]ost of the uncertainty surrounding CL rules stems from the vagueness associated with the reformulations that writers or content developers are expected to make". Daher ist es essentiell, dass die Regeln nach dem gleichen Muster bei allen ihrer analysierten Sätze umgesetzt werden. Wendet der Forscher mehrere Umsetzungsmuster an, so muss er bei der Analyse den Einfluss dieser berücksichtigen. In der vorliegenden Studie gab es z. B. bei der Regel „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden" (Bsp.: Die Herstellerangaben *sind* stets *zu beachten*) zwei mögliche Umsetzungsmöglichkeiten: Imperativ am Satzanfang (*Beachten Sie* stets die Herstellerangaben) oder Imperativ am Satzende (Stets die Herstellerangaben *beachten*). Die Entscheidung fiel auf die erste Umsetzungsmöglichkeit und sie wurde konsequent bei allen Sätzen dieser Regel verwendet, um die Anzahl der unabhängigen Variablen einzuschränken (für mehr dazu siehe §4.5.2.2).

Weitere Herausforderungen bzw. Überlegungen, die bei der MÜ-Evaluation im Kontext der KS berücksichtigt werden sollten, drehen sich hauptsächlich um die *Testpersonen* (ihre *Anzahl* und ihr *Profil*) sowie um die für die Evaluation bzw. Analyse zur Verfügung stehende *Zeit* (Nyberg u. a. 2003: 257). Die wesentlichen Merkmale bei dem Profil der Testpersonen klären die Fragen, ob sie Muttersprachler sind, zu welchem Grad sie mit dem Thema und der KS vertraut sind, ob sie die KS-Texte den unkontrollierten gegenüber bevorzugen (ebd.). Diese Aspekte wurden in der vorliegenden Studie folgendermaßen berücksichtigt: Es wurde ein relativ großer Datensatz (1.100 Sätze) geschaffen, der vollständig von den acht teilnehmenden Übersetzern evaluiert wurde. Bei den Teilnehmern handelte es sich um Semiprofessionals (Jääskeläinen 1993: 99f.). Da professionelle Übersetzer nicht selten eine kritische Einstellung gegenüber der MÜ haben (vgl. O'Brien 2006: 126), würde sich diese störend auf die Analyse des KS-Einflusses auswirken. Sechs Teilnehmer gaben beim Posttest an, dass sie die MÜ in Kombination mit PE verwenden. Dies deutet darauf hin, dass sie die MÜ grundsätzlich nicht ablehnen und mit ihren Schwächen umgehen können. Bezüglich ihrer Vertrautheit mit der KS kamen alle Bewerter mit dem Thema KS im Rahmen ihres Studiums in Kontakt. Nur eine Teilnehmerin hatte kurze berufliche Erfahrung mit der KS. Da die Mehrheit der Teilnehmer keine berufliche Erfahrung mit der KS hatte, war ihre Einstellung zur KS noch nicht gefestigt (weder positiv noch negativ); Näheres zu dem Teilnehmerprofil und seiner Analyse unter §4.5.5.3 bzw. §5.2.5. In Bezug auf den Zeitfaktor wurde die Fehlerannotation ohne Zeitdruck über zwei Monate durchgeführt; Genaueres dazu unter §4.5.3.1 Schritt [7] sowie §4.5.4.2. In der

Humanevaluation wurden die Sätze auf 44 Tests randomisiert aufgeteilt. Jeder Übersetzer hatte die Möglichkeit min. 1 und max. 3 Tests pro Tag zu bewerten. Somit ging die Bewertung für die Teilnehmer mit einer gewissen zeitlichen und örtlichen Freiheit und Flexibilität einher. Der „Ablauf der Evaluation“ ist unter §4.5.5.2 ausführlich präsentiert.

3.6 Fazit

Dieses Kapitel befasste sich mit der maschinellen Übersetzung. Es lieferte eine Beschreibung der Entwicklung maschineller Übersetzungssysteme bis zur Einführung der neuronalen MÜ-Systeme. Es folgten eine Diskussion des Themas MÜ-Qualität, des Designs einer Evaluation sowie eine Darstellung der verschiedenen MÜ-Evaluationsmechanismen. Schließlich wurden zahlreiche vorherige MÜ-Studien im Bereich der KS präsentiert und die Schwierigkeiten einer Untersuchung auf KS-Regelebene thematisiert. Im Zusammenhang mit den früheren MÜ-Ansätzen (RBMÜ, SMÜ und HMÜ) zeigen die bisherigen Studien, dass die KS indirekt zur Verbesserung der (maschinellen) Übersetzbarkeit beiträgt, da sie im Allgemeinen die Ambiguität und die Satzkomplexität reduziert sowie die Satzstruktur vereinfacht. Durch die Diskussion der vorherigen MÜ-Studie im Kontext der KS werden drei Forschungslücken erkennbar: bislang ist die *NMÜ* im Kontext der KS unerforscht; die KS-Regeln der *deutschen* Sprache wurden unzureichend untersucht; die KS-Wirkung auf *Regelebene* wurde ungenügend geprüft. Die vorliegende Studie widmet sich diesen Forschungslücken. Empirisch werden diverse KS-Regeln analysiert, um ihre direkten bzw. indirekten Auswirkungen auf den MÜ-Output sowohl stilistisch als auch inhaltlich im Sinne der Verständlichkeit und Genauigkeit angesichts des jüngsten markanten MÜ-Fortschritts insbesondere bei der Lieferung flüssigen und stilistischen Outputs zu testen. Damit die KS-Auswirkungen im Verlauf der MÜ-Entwicklungsphasen in den letzten Jahren verfolgt und ermittelt werden können, werden in der Analyse die bekanntesten MÜ-Ansätze (RBMÜ, SMÜ, HMÜ und NMÜ) mithilfe humaner sowie automatischer Evaluationen verglichen. Im folgenden Kapitel wird die hierfür angewandte Methodologie aufbauend auf den theoretischen Grundlagen in diesem und dem vorherigen Kapitel im Detail präsentiert.

4 Methodologie

4.1 Einleitung

In diesem Kapitel wird die Forschungsmethodik der Arbeit vorgestellt. Wie das vorherige Kapitel zeigt, existiert für Evaluationsstudien im Bereich der KS in Zusammenhang mit der MÜ keine Standardmethodik. Vor diesem Hintergrund wird in der vorliegenden Studie ein Mixed-Methods-Triangulationsansatz angewandt, auf dessen Basis die Vorteile der Replizierbarkeit eines quantitativen Ansatzes und der Realitätsnähe eines qualitativen Ansatzes ausgeschöpft sowie die Nachteile des jeweiligen Ansatzes minimiert werden konnten. Das Kapitel ist wie folgt strukturiert: Zunächst werden die Forschungsmethodikgefolgt von der Operationalisierung und der Validität der Arbeit genauer erläutert. Das Studiendesign kann als eine Reihe von Entscheidungen betrachtet werden, die der Forscher im Laufe seiner Forschung unter Berücksichtigung der Reliabilitäts- und Validitätsaspekte trifft, nachdem er aus den Fehlern vorheriger Studien gelernt bzw. von ihren Ergebnissen profitiert hat. Daher widmet sich der Abschnitt Studiendesign der ausführlichen Darstellung der Auswahl der analysierten Regeln und der untersuchten MÜ-Systeme wie auch der detaillierten Präsentation des Aufbaus des Datensatzes. Außerdem werden im Rahmen dieses Abschnitts die methodischen Überlegungen und Testläufe, die zum Design der drei implementierten Evaluationsmethoden geführt haben, sowie die Vorgehensweisen der durchgeführten Analysen präsentiert.

4.2 Forschungsmethodik

Wie die Literaturübersicht im Kapitel 3 darlegt, stellt die Bewertung der Qualität des Outputs der maschinellen Übersetzung (MÜ) eine komplexe Aufgabe dar, die noch komplexer wird, wenn sie in Zusammenhang mit der Kontrollierten Sprache untersucht wird. Daher war es erforderlich einen Forschungsansatz anzuwenden, mit dem diese Komplexität behandelt werden kann. Mithilfe eines dreiphasigen Mixed-Methods-Triangulationsansatzes strebt die Studie an, den

Einfluss einzelner Regeln der Kontrollierten Sprache auf die Qualität[1] des MÜ-Outputs zu untersuchen. Nachfolgend wird der Mixed-Methods-Ansatz, seine Umsetzung über drei Phasen sowie die Triangulation der angewandten Methoden näher erläutert.

Eine der ersten Definitionen des Mixed-Methods-Ansatzes entstand 1989 von Greene, Caracelli und Graham im Bereich der Evaluation. Nach dieser Definition umfasst ein Mixed-Methods-Design mindestens eine quantitative Methode („designed to collect numbers") und eine qualitative Methode („designed to collect words") (Creswell & Clark 2007: 2). Johnson & Onwuegbuzie (2004) definierten den Mixed-Methods-Ansatz als „the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study". Im Grunde wurde das Konzept des Mixed-Methods-Ansatzes auf verschiedene Weise definiert. Vor diesem Hintergrund untersuchten Johnson u. a. (2007) in einem häufig zitierten Artikel des Journal of Mixed Methods Research (JMMR) 19 Definitionen sehr bekannter Studien. Der Artikel reflektiert die zahlreichen Unterschiede, die diese Studien, unter anderem in Bezug auf die Faktoren „was wird gemixt" (z. B. Methoden, Methodologien, Forschungstypen), „wann findet das Mixing statt" (z. B. bei der Datenerhebung, -analyse, -interpretation) sowie „Ziel des Mixing" aufweisen. Sie zeigen, dass das Wort „Methods" (in „Mixed Methods") eine breite Interpretation und Verwendung hat und somit die Einbeziehung von Fragestellungen und Strategien hinsichtlich der Methoden der Datenerhebung sowie der Forschungsmethoden zulässt (Johnson u. a. 2007). Im Endeffekt besteht die zentrale Prämisse dieses Ansatzes darin, dass "the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone" (Creswell & Clark 2007: 5). Bei der vorliegenden Studie ging es nicht nur darum, durch den Mixed-Methods-Ansatz ein besseres Verständnis der Forschungsprobleme zu erlangen, vielmehr wäre das Ziel der Studie ohne diesen Ansatz nicht erreichbar. Denn jede der implementierten Methoden lieferte Daten, die für die Analyse in der darauffolgenden Methode erforderlich war, wie unten näher beschrieben ist.

Des Weiteren wurden die Ergebnisse der drei angewandten Methoden (Fehlerannotation, Humanevaluation und automatische Evaluation des MÜ-Outputs) trianguliert. Der Triangulationsansatz wurde in den 1970er Jahren von Denzin entworfen und als die „combination of methodologies in the study of the same phenomenon" definiert (Kuckartz 2014: 44f.). Mit der Kombination mehrerer Methoden zielt man darauf ab, das Vertrauen in die Validität der Ergebnisse zu stei-

[1] Für die Definition der Qualität im Rahmen dieser Studie siehe §4.5.5.1.

gern (Frey u. a. 1991: 24; Kuckartz 2014: 47). Saldanha & O'Brien (2014: 23) betrachten die methodologische Triangulation als „the backbone of solid, high quality research". Bei einer Methodentriangulation sind drei Resultate möglich (Erzberger & Kelle 2003): eine vollständige oder Teilübereinstimmung der Ergebnisse, sich gegenseitig ergänzende oder unterschiedliche bzw. widersprüchliche Ergebnisse. Alle genannten Arten der Ergebnisse können aufschlussreich sein: eine Übereinstimmung untermauert die Ergebnisse; sich ergänzende Resultate können das Ergebnisbild vervollständigen; und unterschiedliche bzw. widersprüchliche Ergebnisse können die Aufmerksamkeit des Forschers auf eine offene Frage richten.

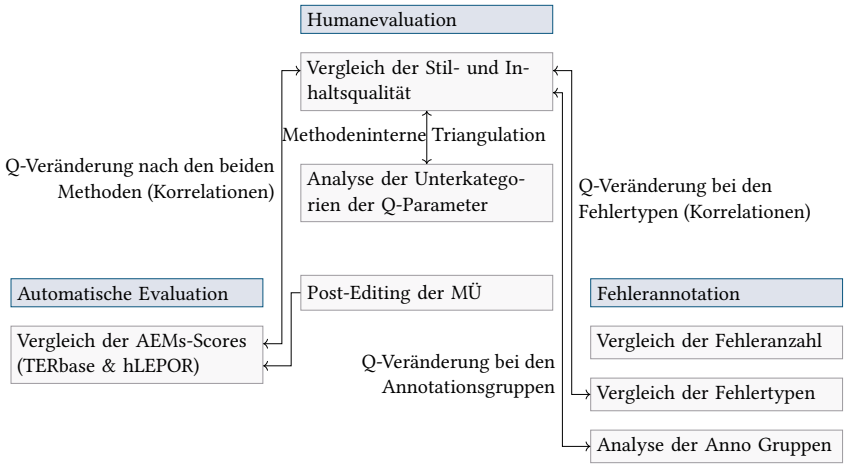


Abbildung 4.1: Methodik der Arbeit – Mixed-Methods-Triangulationsansatz

Die quantitativen und qualitativen Daten können gleichzeitig oder in aufeinanderfolgenden Phasen erhoben und analysiert werden (Johnson & Onwuegbuzie 2004; Saldanha & O'Brien 2014: 23). In der vorliegenden Studie wurden die drei genannten Methoden aufeinander aufgebaut und entsprechend in der unten dargestellten Reihenfolge durchgeführt (Abbildung 4.1).

Grundsätzlich bildet der Datensatz zwei Szenarien („vor" vs. „nach" dem Einsatz der KS-Regel) ab, die mithilfe der drei Methoden bottom-up analysiert und verglichen werden. In der ersten Phase wurden quantitative Daten aus der Fehlerannotation (siehe §4.5.4.3) gewonnen, die durch den Vergleich der beiden Szenarien eine Abnahme bzw. Zunahme der Fehleranzahl aufzeigen und Aufschluss über die aufgetretenen Fehlertypen geben. Ferner wurden die aufgetretenen Fehlerty-

pen qualitativ analysiert. Darüber hinaus wurden Annotationsgruppen[2] gebildet, nach denen zwischen fehlerfreien und fehlerhaften MÜ vor und nach dem Einsatz der KS-Regel unterschieden wird. Diese quantitativen und qualitativen Analysen der Fehlerannotation zeigen zwar eine Tendenz eines potenziellen Anstiegs bzw. Abstiegs der Qualität an, jedoch kann auf deren Basis keine klare Aussage zu der Qualität getroffen werden. Dies konnte erst in der zweiten Phase durch die Humanevaluation (siehe §4.5.5.4) anhand der quantitativen Daten (Qualitätsbewertung auf Likert-Skalen in den beiden Szenarien) sowie weiterer qualitativer Daten (Angabe und Erläuterung der Qualitätskriterien) aufgedeckt werden. Denn die Humanevaluation lieferte zum einen eine direkte Bewertung der Qualität des MÜ-Outputs und zum anderen konnte der Zusammenhang zwischen den Fehlertypen (aus der ersten Phase) und der Qualität quantitativ ermittelt und qualitativ anhand der Erläuterungen der Bewerter näher beleuchtet werden. Ein weiterer Output der zweiten Phase stellte eine Alternativübersetzung dar, die die Übersetzer für jede stilistisch bzw. inhaltlich kritisierte oder fehlerhafte MÜ vorschlugen. Die dritte Phase baut weiter auf der zweiten Phase auf, indem die vorgeschlagenen Alternativübersetzungen zur Qualitätsbewertung mittels zwei automatischen Evaluationsmetriken „TERbase“ und „hLEPOR“ (vor vs. nach dem Einsatz der KS-Regel) herangezogen wurden (siehe §4.5.6.5). Darüber hinaus wurde die Korrelation zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der Qualität in den beiden Szenarien untersucht, um herauszufinden, inwiefern die Ergebnisse der beiden Methoden übereinstimmen (d. h. ob ein Anstieg der Qualität gleichzeitig mit besseren AEM-Scores vorkam bzw. ein Abstieg der Qualität mit schlechteren AEM-Scores einherging). Auf diese Weise wurde der Mixed-Methods-Ansatz in dieser Studie eingesetzt, um die Forschungsfrage behandeln zu können.

Denzin unterscheidet zwischen einer methodeninternen Triangulation („within method triangulation“) und einer methodenexternen Triangulation („between method triangulation“) (Kuckartz 2014: 47). Neben der methodenexternen Triangulation zwischen den drei genannten Methoden wurde bei der Humanevaluation eine methodeninterne Triangulation angewendet (Abbildung 4.6). Hierbei bestand die Aufgabe[3] darin, für die Übersetzungsqualität eine Punktzahl auf

[2] Die Daten wurden binär bzw. dichotom aufgeteilt (keine Fehler aufgetreten „0“; Fehler aufgetreten „1“), daraus wurden vier Annotationsgruppen in Bezug auf die KS-Stelle gebildet: (1) RR: MÜ ist vor und nach dem Einsatz der KS-Regel fehlerfrei; (2) FF: MÜ beinhaltet vor und nach dem Einsatz der KS-Regel Fehler; (3) RF: MÜ ist nur vor dem Einsatz der KS-Regel fehlerfrei; (4) FR: MÜ ist nur nach dem Einsatz der KS-Regel fehlerfrei. Für die Definition der KS-Stelle siehe §4.5.2.1.

[3] Für eine genaue Beschreibung der „Testaufgaben“ siehe „Darstellung der Evaluation“ unter §4.5.5.2.

der Likert-Skala zu vergeben, nachdem der Teilnehmer die relevanten Qualitätskriterien angekreuzt und kurz kommentiert bzw. die Alternativübersetzung vorgeschlagen hatte. Durch den Einsatz von mehreren Techniken innerhalb der Humanevaluation konnte die interne Konsistenz erhalten, die Zuverlässigkeit optimiert und die Daten genauer interpretiert werden.

Zusammenfassend zeigt die obige kurze Darstellung der angewandten Methoden, wie die durchgeführten qualitativen und quantitativen Analysen sich bei der Erarbeitung der Forschungsfragen gegenseitig ergänzten. Anhand der qualitativen Daten und Analysen konnten die quantitativen Daten erörtert bzw. interpretiert werden. Gleichzeitig wurden die quantitativen Daten auf Basis der statistischen Tests validiert. Auf diese Weise verleihen die quantitativen Analysen der Studie die objektive Validität, die nur mit qualitativen Analysen nicht gegeben wäre. Zudem ermöglichte eine Triangulation der verschiedenen Methoden die Ergebnisse zu untermauern und gleichzeitig Widersprüche zu einer weiteren Untersuchung aufzudecken.

4.3 Operationalisierung

Durch die Operationalisierung beschreibt der Forscher die beobachtbaren Merkmale des untersuchten Konzepts (Frey u. a. 1991: 95). Genau genommen bezieht sich die Operationalisierung auf „the operations involved in measuring the dependent variable“ (Saldanha & O’Brien 2014: 24). Bei der Operationalisierung müssen drei Kriterien erfüllt sein (Frey u. a. 1991: 95): *Adäquatheit* durch eine umfassende Beschreibung des untersuchten Konzepts, *Exaktheit* i. S. v. allgemein anerkannt und *Klarheit* für den Leser. Die Operationalisierung betrifft nicht nur quantitative Forschungsansätze, sondern ist gleichermaßen bedeutsam für qualitative Ansätze (Saldanha & O’Brien 2014: 24).

Als abhängige Variable steht die MÜ-Qualität in der vorliegenden evaluativen Studie im Mittelpunkt. Sie wird umfassend auf Basis dreier bewährten Evaluationsmethoden anhand einer Reihe von klar definierten quantitativen und qualitativen Messwerten, die im Bereich der MÜ-Evaluation allgemein anerkannt sind, gemessen. Der Einsatz einer Fehlertypologie ist ein herkömmlicher Ansatz zur Qualitätsbewertung, nicht nur bei der MÜ, sondern ursprünglich auch in der Translationswissenschaft (vgl. Vilar u. a. 2006; Saldanha & O’Brien 2014: 101). Die Humanevaluation sowie die automatische Evaluation sind die klassischen Methoden der Qualitätsbewertung in der MÜ (siehe §3.4.3.1 und §3.4.3.2). Die Stil- und Inhaltsqualität wurden in Anlehnung an Hutchins und Somers definiert (siehe §4.5.5.1). Der Vergleich der Qualität des MÜ-Outputs vor vs. nach dem Einsatz

der einzelnen KS-Regeln erfolgte auf Grundlage der nachstehenden qualitativen und quantitativen Daten (Tabelle 4.1):

Tabelle 4.1: Überblick der analysierten Daten vor und nach dem KS-Einsatz bei den angewandten Methoden

Messwerte	**Evaluationsmethoden**
• Vergleich der Fehleranzahl im Allgemeinen • Vergleich der Fehlertypen • Vergleich der Fehleranzahl jedes Fehlertyps • Analyse der Aufteilung der Annotationsgruppen	auf Basis der Fehlerannotation
• Vergleich der Stil- und Inhaltsqualität (5er-Likert-Skala) • Analyse der beeinflussten Qualitätskriterien (Checkboxen) • Analyse der Korrelation zwischen den Fehlertypen und der Stil- und Inhaltsqualität • Analyse der Qualität jeder Annotationsgruppe	auf Basis der Humanevaluation
• Vergleich der AEM-Scores • Analyse der Korrelation zwischen den AEM-Scores und der Stil- und Inhaltsqualität	auf Basis der automatischen Evaluation

Nach dieser Darstellung der Forschungsmethodik und der Operationalisierung richten wir nachfolgend den Blick auf die Validität, das Studiendesign sowie seine Details bei den einzelnen Evaluationsmethoden.

4.4 Validität

Durch die Validität strebt der Forscher an, die Genauigkeit seiner Schlussfolgerungen (sog. Interne Validität) und die Generalisierbarkeit seiner Ergebnisse (sog. Externe Validität) sicherzustellen (Frey u. a. 1991). Die *interne Validität* kann durch drei Quellen gefährdet werden: durch den Forscher (d. h. durch seinen „personal attribute effect“ bzw. „unintentional expectancy effect“), die Studienvorgehensweise (i. S. v. die angewandten Verfahren im Allgemeinen und während der Datenanalyse) und die Studienteilnehmer (bzgl. ihrer Auswahl, ihres Verhaltens und der Entwicklung dieses Verhaltens im Laufe der Studie sowie bzgl. eines potenziellen gegenseitigen Einflusses). Die *externe Validität* kann von drei Faktoren

beeinflusst werden: dem Sampling, der ökologischen Validität und der Forderung nach einer Replikation. (ebd.: 125ff.) Im Folgenden wird erläutert, wie die genannten Aspekte der internen und externen Validität in der Studie bei der Erstellung des Datensatzes sowie der Durchführung der Analysen berücksichtigt wurden.

Für die Erstellung des Datensatzes wurde ein Korpus aus zehn technischen Dokumenten verschiedener Unternehmen erstellt. Die Ausgangssätze, die einen Verstoß gegen eine/mehrere der analysierten KS-Regeln aufwiesen, wurden automatisch mithilfe des CL-Checkers CLAT[4] (Rösener 2010) identifiziert. Aus den identifizierten Sätzen wählte die Forscherin die Testsätze nach klar definierten Kriterien unter Berücksichtigung einer möglichst ausgewogenen Aufteilung aus allen Quellen aus. Die Aufbereitung der Ausgangssätze und der Einsatz der KS-Regeln wurde von der Forscherin nach einem vordefinierten Muster durchgeführt. Es folgte eine Qualitätssicherung, in der über zwei Schritte die Qualität der Ausgangssätze vor und nach dem Einsatz der KS-Regeln orthografisch, grammatisch und stilistisch von einer erfahrenen Übersetzerin und zwei professionellen Linguisten geprüft wurde. Wurde ein Satz von den Prüfern kritisiert, so wurde er durch einen neuen Satz ersetzt, der ebenfalls geprüft wurde. Für die genaueren Details zur Erstellung und Aufbereitung des Datensatzes siehe §4.5.3.1.

Bezüglich der Validität der Selektion der untersuchten Regeln wurden diese nach vordefinierten Auswahlkriterien (siehe §4.5.2.1) festgelegt. Nicht alle analysierten Regeln haben nach tekom (2013) die Übersetzbarkeit im Fokus; einige Regeln zielen auf die Lesbarkeit bzw. die Verständlichkeit ab. Jedoch zeigen die weiteren zitierten Studien (siehe §4.5.2.2), dass die analysierten Regeln durch die Reduzierung der Satzkomplexität, Vereinfachung der Satzstruktur bzw. Verminderung der Ambiguität die maschinelle Übersetzbarkeit – bei den früheren MÜ-Ansätzen – verbessern (vgl. Bernth & Gdaniec 2001; Reuther 2003; Siegel 2011; Congree 2018). Auf diesem Wege tragen auch Regeln, die die Verständlichkeit in den Mittelpunkt stellen, indirekt zur Verbesserung der Übersetzbarkeit bei. Diese allgemeine Wirkung der KS-Regeln auf die MÜ wurde von Fiederer & O'Brien (2009: 53) bestätigt und wird im Rahmen dieser Studie angesichts der Entwicklung des jüngsten MÜ-Ansatzes der NMÜ und seines vielversprechenden flüssigen Outputs getestet.

Die Fehlerannotation wurde von einem in DE-EN vereidigten Übersetzer mit sechs Jahren Berufserfahrung durchgeführt. Anschließend wurden die annotierten Fehler von zwei professionellen Linguisten geprüft. Aufgrund der großen Anzahl der MÜ-Sätze (2.160 Sätze) prüfte jeder Linguist die Hälfte der MÜ-Sätze. Im Fall, dass er der Annotation nicht zustimmte, musste er die Übersetzung ebenfalls annotieren. Der zweite Linguist prüfte im Anschluss beide Annotationen

[4] http://www.iai-sb.de/de/produkte/clat [abgerufen am 23.12.2014]

und entschied sich für eine davon. Für die Details der Fehlerannotation siehe §4.5.4.

In der Humanevaluation wurden die Teilnehmer nach einem „purposive sampling" ausgewählt (vgl. O'Brien 2006: 127), d. h. sie mussten bestimmte Kriterien in Bezug auf die Sprachkompetenzen und die translatorischen Qualifikationen (siehe §4.5.5.3) erfüllen. Die Evaluation wurde von einer diversifizierten Teilnehmergruppe von acht (vier weiblichen und vier männlichen) qualifizierten Übersetzern mit Englisch als Muttersprache aus zwei Ländern (zwei aus England und sechs aus den USA) durchgeführt. Gleichzeitig war die Teilnehmergruppe weitgehend homogen in Bezug auf die Teilnehmerqualifikation und -erfahrung. Alle Teilnehmer sind semiprofessionelle Übersetzer (Jääskeläinen 1993: 99f.). Sie besaßen bereits einen Bachelorabschluss im Fach Translation und befanden sich im letzten Semester des Masterstudiengangs Translation. Die Übersetzungserfahrung von sieben Teilnehmern variierte zwischen einem und zwei Jahren. Die letzte Teilnehmerin hatte vier Jahre Erfahrung als Übersetzerin in Teilzeit. Ursprünglich fiel die Entscheidung zwischen professionellen und semiprofessionellen Übersetzern aufgrund der eingeschränkten Verfügbarkeit und der hohen Kosten von professionellen Übersetzern auf semiprofessionelle. Im Nachhinein zeigte der Post-Test, dass die Einstellung der Teilnehmer zur MÜ grundsätzlich als neutral eingestuft werden kann: Sechs Teilnehmer verwenden MÜ-Systeme. Die übrigen zwei beschäftigen sich überwiegend mit literarischen Übersetzungen und verwenden die MÜ selten. Professionelle Übersetzer haben nicht selten eine kritische Einstellung zur MÜ (vgl. O'Brien 2006: 126). Da die Studie sich nicht mit der MÜ-Qualität an sich beschäftigt, sondern gezielt mit dem Effekt der KS auf die MÜ, könnte eine allgemein negative Einstellung zur MÜ für die Analyse nicht zielführend sein bzw. sich negativ auf das Ergebnis auswirken. Somit erwies sich die Entscheidung, semiprofessionelle Übersetzer einzustellen, als sinnvoll. Ferner wurde zur Förderung der Teilnehmermotivation die Teilnahme vergütet.

Alle Interessenten hatten die gleiche Chance an der Studie teilzunehmen. Sobald ein Interessent, der die Auswahlkriterien erfüllte, die Forscherin kontaktierte, wurde er in die Studie aufgenommen. Die Qualitätsdefinitionen wurden klar dargestellt und die Testanweisungen wurden mithilfe von Beispielen erläutert (siehe Anhang A). Alle Teilnehmer erhielten zuerst einen Probetest, mit dessen Hilfe sichergestellt wurde, dass sie mit dem Testablauf, -aufbau bzw. den Testanweisungen vertraut sind. Alle Rückfragen wurden vor Beginn der Testphase ohne jeglichen Hinweis zum Ziel und Hintergrund der Studie geklärt. Jeder Teilnehmer durfte den Test zu einer für ihn passenden Uhrzeit und ortsflexibel durchführen, sofern er konzentriert und ohne Unterbrechung arbeiten konnte. Diese

Flexibilität bzw. die Durchführung ohne Anbindung der Teilnehmer an einen fremden Ort mit vorgegebenen kontrollierten Settings trägt zur Erhöhung der ökologischen Validität bei.

Eine Teilnahmevoraussetzung bestand darin, dass der Teilnehmer mindestens einen Test pro Tag durchführt, um eine Unterbrechung zu vermeiden, die ggf. sein Verhalten und somit das Intrarater-Agreement hätte beeinflussen können. Dadurch, dass die Evaluation während der Semesterferien stattfand, waren die Teilnehmer nicht an der Universität. Die Mehrheit von ihnen war in ihrer Heimat. Dies schließt entsprechend das Risiko eines „intersubject bias" weitgehend aus (vgl. O'Brien 2006: 127), d. h. die Möglichkeit, dass die Teilnehmer sich während der Testphase austauschen und entsprechend gegenseitig beeinflussen.

Die MÜ-Sätze aller Systeme vor und nach dem KS-Einsatz wurden randomisiert auf 44 Tests aufgeteilt. Zudem erhielten die Teilnehmer die Tests in unterschiedlicher Reihenfolge. Diese doppelte Randomisierung trug dazu bei, zwei Risiken zu minimieren: erstens, das Risiko einer „Maturation", d. h. „sentences they graded later in the cycle will get a different look than the ones they graded earlier" (White 2003: 209); zweitens, dass Qualitätsbewertungen zwischen benachbarten Sätzen so unabhängig wie möglich gehalten werden (vgl. Hamon 2007).[5] Die Tests wurden in Form von Excel-Tabellen erstellt (Abbildung 4.7), einem Programm, mit dem alle Teilnehmer vertraut sind. Dies unterstützt die ökologische Validität. Zur Sicherstellung der Vollständigkeit der Daten war der Testablauf so gestaltet, dass der Teilnehmer am Ende des Tages die bewerteten Tests abgibt. Sie wurden sofort auf Vollständigkeit geprüft. Anschließend erhielt er die Tests des folgenden Tages.

Die Qualitätsdefinitionen (Hutchins & Somers 1992, siehe §4.5.5.1) wurden in den Qualitätskriterien integriert (Abschnitt [3a] und [3b] in Abbildung 4.6), d. h. nicht nur in Form von Definitionen am Anfang des Tests zur Verfügung gestellt, wie es in Evaluationsstudien typisch ist. Die Teilnehmer wurden aufgefordert, die zutreffenden Qualitätskriterien anzukreuzen, zu kommentieren bzw. posteditieren und anschließend einen Score zu vergeben. Somit hatten alle Teilnehmer eine direkte und einheitliche Basis für die Vergabe der Qualität-Scores, die sicherlich zu den hohen Intrarater- und Interrater-Agreements beiträgt.

Ferner wurden die Ausgangssätze den Teilnehmern bei der Evaluation zur Verfügung gestellt. Da der Test zum Teil eine PE-Aufgabe beinhaltet und die Norm im PE lautet, dass dem Post-Editor der Ausgangstext zur Verfügung steht (O'Brien 2006: 130), unterstützt die Darstellung der Ausgangssätze die ökologische Validität (ebd.).

[5] Beide Risiken sind ausführlich unter „Ablauf der Evaluation" (§4.5.5.2) erläutert.

Da der Fokus der Studie darin liegt, einen potenziellen Einfluss der einzelnen KS-Regeln zu untersuchen, war es als erstes erforderlich, die Stelle in der MÜ zu ermitteln, die durch den Einsatz der KS-Regel beeinflusst werden konnte (bezeichnet als die „KS-Stelle“).[6] Fehler in der MÜ, die außerhalb der KS-Stelle lagen, mussten vor der Humanevaluation korrigiert werden, um sicherzustellen, dass der Unterschied in den beiden Versionen „vor KS“ und „nach KS“ nur an der KS-Stelle auftritt. Dieser Schritt wurde von der Forscherin durchgeführt. Anschließend wurde die Qualität der MÜ-Sätze nach der Korrektur der Fehler außerhalb der KS-Stelle über zwei Phasen von drei qualifizierten erfahrenen Übersetzern auf potenzielle Fehler sowie stilistische Akzeptanz außerhalb der KS-Stelle geprüft. Wurde eine MÜ kritisiert, so wurde sie von der Humanevaluation ausgeschlossen (Genaueres dazu unter §4.5.3.1 Schritt [10]).

Ferner wurde darauf geachtet, dass die MÜ-Sätze in Bezug auf die Existenz bzw. Nicht-Existenz von Fehlern in der Humanevaluation in ähnlichem Verhältnis wie in der Annotation repräsentiert sind. Dieser Schritt ist für die Datenanalyse essentiell, da eine Unausgewogenheit zur Verzerrung der Ergebnisse der Humanevaluation führen würde. Die Ausgewogenheit der MÜ-Sätze wurde nach definierten Kriterien realisiert (siehe §4.5.3.2).

In Bezug auf den unterschiedlichen Prozentsatz der analysierten MÜ-Sätze aus jedem System (Abbildung 4.3): Da der NMÜ-Output eine sehr geringe Anzahl von Fehlern enthielt und die Auswahlkriterien für MÜ-Sätze diejenigen Sätze für die Humanevaluation ausschlossen, die eine hohe Anzahl von Fehlern enthielten (siehe Kriterien in [8] unter §4.5.3.1), wurde eine größere Anzahl von NMÜ-Sätzen im Vergleich zu den Sätzen anderer Systeme ausgewertet. Obwohl die Bewertung einer gleichen Anzahl von Sätzen jedes MÜ-Systems aus Gründen der Unverzerrtheit als ideal betrachtet werden kann, könnte dies nur durch eine Reduktion der Anzahl der analysierten NMÜ-Sätze realisiert werden. Für dieses Vorgehen wäre es wichtig gewesen, konkrete Ausschlusskriterien für die Reduzierung der Anzahl von NMÜ-Sätzen zu definieren. Bei solchen Ausschlusskriterien wären drei Szenarien denkbar: (1) Ausschluss fehlerhafter NMÜ-Sätze, was zu deutlich besseren Ergebnissen bei Google Translate als die aktuellen Ergebnisse führen würde; (2) Ausschluss fehlerfreier NMÜ-Sätze. Letzteres wäre ein subjektives Vorgehen, da es mit den anderen Systemen nicht realisierbar wäre. Das letzte mögliche Szenario wäre ein (3) Ausschluss einer Mischung aus fehlerfreien und fehlerhaften NMÜ-Sätzen. Das würde jedoch erneut die Frage nach den anzuwendenden Ausschlusskriterien aufwerfen. Aus diesem Grund war der Prozentsatz der analysierten NMÜ-Sätze ein wenig höher als der der anderen Systeme.

[6] In der Studie bezeichnet die „KS-Stelle“ den Teil des Ausgangssatzes, der bei dem Einsatz der KS-Regel modifiziert werden muss, und sein Äquivalent im Zielsatz.

Die automatische Evaluation wurde zur Erhöhung der Objektivität mithilfe zweier Metriken gemessen. Basis der Messung war die Verwendung von Referenzübersetzungen, die von den Teilnehmern der Humanevaluation angegeben wurden. Zur Berücksichtigung von unterschiedlichen PE-Möglichkeiten wurden zwei Referenzübersetzungen pro MÜ-Satz verwendet. Bei der Auswahl der zwei Referenzübersetzungen wurde darauf geachtet, dass die Referenzübersetzungen möglichst von allen Teilnehmern gleichermaßen verwendet werden (§4.5.6.2).

Wie von Frey u. a. (1991) empfohlen, wurden mehrere Testläufe (Pilottests) zur Reduzierung von Messfehlern durchgeführt. Konkret wurden Testläufe bei der Entscheidung zur Darstellung/Nichtdarstellung der Ausgangssätze und Markierung/Nichtmarkierung der KS-Stelle während der Evaluation, Prüfung der Klarheit der Testanweisungen bzw. -aufgaben, Festlegung der Anzahl der Sätze pro Test sowie Qualitätssicherung der Ausgangs- und Zielsätze durchgeführt. In Bezug auf die Objektivität umfassten die drei implementierten Methoden eine große Anzahl an quantitativen Messungen. Bei den Analysen wurden die Daten systematisch und regelmäßig in definierten Excel-Tabellen aufgezeichnet und innerhalb kurzer Zeit dokumentiert. Gleichzeitig wurden die quantitativen Messungen durch die Triangulation – wie in den vorherigen Abschnitten dargestellt – mit qualitativen Analysen kombiniert. Die Triangulation erhöht wiederum die Messvalidität (measurement validity) (ebd.: 124). Um eine Replikation zu ermöglichen, strebte die Forscherin danach, die Schritte der Datenerhebung, -aufbereitung, -analyse sowie die methodische Vorgehensweise zusammen mit der Begründung sämtlicher getroffenen Entscheidungen im folgenden Abschnitt ausführlich zu dokumentieren.

4.5 Studiendesign

4.5.1 Auswahl des analysierten Sprachenpaars und der MÜ-Systeme

Das Ziel der Studie besteht darin, den Einfluss des Einsatzes einzelner KS-Regeln auf die Qualität des MÜ-Outputs bei verschiedenen MÜ-Ansätzen zu untersuchen. Die vier bekannten MÜ-Ansätze (RBMÜ, SMÜ, HMÜ und NMÜ) wurden durch die Analyse von fünf MÜ-Systemen analysiert. Somit hat die Studie drei Dimensionen: die KS-Regeln, die MÜ-Systeme und das Sprachenpaar. Für den einzelnen Vergleich der neuen Regeln und der fünf Systeme wurde die Sprachenpaarvariable auf Deutsch > Englisch fixiert. Das analysierte Sprachenpaar ist für international agierende Unternehmen aus deutschsprachigen Ländern von großer Relevanz. Innerhalb der Europäischen Union fordert die Maschinenrichtlinie 2006/42/EG, dass alle schriftlichen und verbalen Informationen und Warnhinweise in der Amtssprache des EU-Mitgliedstaates beiliegen müssen, in der das

Produkt in Verkehr gebracht bzw. in Betrieb genommen wird (tekom RG Alb Donau 2010). Dementsprechend ist eine englische Übersetzung der Betriebsanleitung für Irland, England und Malta erforderlich. Des Weiteren stellen weitere englischsprachige Länder außerhalb der EU große Märkte für Produkte aus dem deutschsprachigen Raum dar; z. B. waren die USA laut einer aktuellen Statistik nach Exportwert der wichtigste Handelspartner Deutschlands im Jahr 2018 (Statista 2019). Mit diesen Dimensionen und anhand eines angemessen großen Datensatzes war der Umfang der Studie gut zu bewältigen. Eine Erweiterung der Studie auf ein zweites Sprachenpaar wäre aus Zeit- und Kostengründen schwer realisierbar. Die Untersuchung von zwei Sprachenpaaren mit kleineren Datensätzen könnte für die Durchführung einer zuverlässigen statistischen Analyse nicht ausreichend sein. Die Untersuchung weiterer Sprachenpaare ist somit erstrebenswert (siehe §7.3).

Die fünf untersuchten MÜ-Systeme bestanden aus einem RBMÜ-System, einem SMÜ-System, zwei Hybridsystemen sowie einem NMÜ-System. Die Auswahlkriterien der fünf MÜ-Systeme waren, dass sie (1) vom Deutschen ins Englischen übersetzen, (2) kostenlos zugänglich sind und (3) alle MÜ-Ansätze abdecken. Die empirische Studie wurde Ende 2016 durchgeführt. Zu dieser Zeit war Google Translate das erste bzw. einzige online zugängliche NMÜ-System. Da Google Translate ein generisches Black-Box-System ist, mussten alle anderen MÜ-Systeme, um eine einheitliche Untersuchungsbasis zu gewährleisten, ebenfalls generische Black-Box-Systeme sein.[7] Ein Black-Box-System wird wie folgt definiert: „a system which has been trained and tuned a priori and for which we cannot access the model parameters or training data for fine-tuning or improvements" (Mehta u. a. 2020: 1).[8] Da die Hybridsysteme unterschiedlich aufgebaut werden, wurden zwei Hybridsysteme in der Studie verwendet. Folgende MÜ-Systeme wurden verwendet:[9]

- Das regelbasierte System *Lucy LT KWIK Translator* von Lucy Software and

[7] Für den Umgang mit den spezifischen Termini im Rahmen der Studie siehe Schritt [4] unter §4.5.3.1.

[8] Mehr zum Thema Black-Box- vs. Glas-Box-Evaluation unter §3.4.2.

[9] Die obige Beschreibung der Systeme schildert die Systemarchitekturen gemäß ihrem Stand zum Zeitpunkt der Durchführung der Studie (Ende 2016 – Anfang 2017). In der Zwischenzeit wurden die Systeme weiterentwickelt. 2017 bzw. 2018 kamen die neuronalen Systeme von Systran, SDL und Bing auf den Markt: http://www.systransoft.com/systran/translation-technology/pure-neural-machine-translation; https://www.sdl.com/de/about/news-media/press/2018/sdls-neural-machine-translation-sets-new-industry-standards-with-state-of-the-art-dictionary-and-image-translation-features.html; https://www.microsoft.com/dede/translator/blog/2018/11/14/nextgennmt [abgerufen am 16.04.2019]

Services GmbH.[10] Lucy LT ist das Nachfolgesystem des alten METAL MT Systems (Martin & Serra 2014). Nach Alonso und Thurmair ist Lucy LT „a commercial rule-based machine translation system with sophisticated hand-written transfer and generation rules“ (Avramidis u. a. 2014).

- Das statistische System *SDL Free Translation*.[11] Laut SDL arbeitet das System nach einem rein statistischen Ansatz.[12]
- Die Hybridsysteme *Bing* von Microsoft[13] und *Systran*.[14] Bing ist ein „statistisches MÜ-System mit sprachspezifischen Regelkomponenten für das Zerlegen und Zusammensetzen von Sätzen“ (Werthmann & Witt 2014: 84), wobei Systran ursprünglich ein regelbasiertes System war und in den letzten Jahren zu einem hybriden System weiterentwickelt wurde (ebd.).
- Das neuronale System *Google Translate*.[15] Die Übersetzung auf Basis von neuronalen Netzen ist der aktuellste Ansatz der MÜ, den Google Ende 2016 mit der Umstellung seines Systems in Betrieb setzte. Das Modell von Google Translate besteht aus drei Komponenten (Wu u. a. 2016): einem Encoder-Netzwerk, einem Decoder-Netzwerk sowie einem Attention-Netzwerk.[16]

Der Datensatz wurde mit den fünf genannten Systemen übersetzt. Die genaue Vorgehensweise ist unter §4.5.3 erläutert.

4.5.2 Die analysierten KS-Regeln

In diesem Abschnitt wird die Auswahl der analysierten Regeln begründet und die KS-Stelle ausführlich erklärt. Daraufhin folgt eine Darstellung der analysierten Regeln zusammen mit deren Anwendungsbegründungen, Umsetzungsmustern und KS-Stellen. Zum Schluss werden die analysierten Regeln und ihre gezielten Wirkungen im Hinblick auf das untersuchte Sprachenpaar (Deutsch-Englisch) diskutiert.

[10] online zugänglich unter: http://www.lucysoftware.com/english/machine-translation/lucy-lt-kwik-translator-

[11] online zugänglich unter: https://www.freetranslation.com/de

[12] „What Is Machine Translation and How Does It Work?“ https://sdl.uservoice.com/knowledgebase/articles/256030-what-is-machine-translation-and-how-does-it-work. [abgerufen am 06.12.2016].

[13] online zugänglich unter: https://www.bing.com/translator

[14] online zugänglich unter: http://www.systranet.com/translate

[15] online zugänglich unter: https://translate.google.de.

[16] Mehr Details unter §3.3.4.

4.5.2.1 Auswahl der analysierten KS-Regeln

Die in der vorliegenden Arbeit analysierten KS-Regeln stammen aus der Leitlinie „Regelbasiertes Schreiben. Deutsch für die Technische Kommunikation“ (2013) der Gesellschaft für Technische Kommunikation – tekom e. V.[17] Dank einer engen Zusammenarbeit zwischen Experten aus dem Hochschulwesen, der Industrie, Dienstleistungsunternehmen sowie Softwarefirmen bieten die tekom-Regeln ein umfassendes Regelwerk auf sämtlichen Sprach- und Dokumentationsebenen (siehe Tabelle 2.1). Die tekom-Leitlinie stellt einen branchenübergreifenden Standard für die technische Dokumentation dar. Im Mittelpunkt steht die Steigerung der Qualität und die Reduzierung der Kosten im Dokumentations- und Übersetzungsprozess (Drewer & Ziegler 2014: 217f.). Die tekom-Regeln werden aktuell sowohl in der Forschung als auch in der Industrie weitestgehend umgesetzt.[18] Aus diesen Gründen wurden in der Studie Regeln aus dieser Leitlinie untersucht. Die analysierten Regeln wurden gemäß den folgenden Kriterien ausgewählt:

(1) *Regeln, die auf genau einen Satz angewendet werden können:*

Da die Studie den Einfluss einzelner KS-Regeln untersucht, erfolgten die Analysen auf Satzebene. Daher wurden die Regeln, die mehrere Sätze oder die Dokumentstruktur betreffen, ausgeschlossen.

(2) *Regeln, die in allen relevanten Sätzen nach einem festen Muster angewendet werden können:*

Da die Anwendung von verschiedenen Mustern bei dem Einsatz einer KS-Regel das Ergebnis unterschiedlich beeinflussen kann (vgl. Roturier 2006: 74), wurde in allen relevanten Sätzen ein festes Muster angewendet. Dies ermöglichte wiederum, die Anzahl der unabhängigen Variablen in Grenzen zu halten. Die angewendeten Umsetzungsmuster bei den einzelnen Regeln sind in §4.5.2.2 dargestellt.

(3) *Regeln, für die eine KS-Stelle definiert werden kann:*

Zur Untersuchung des Einflusses der einzelnen KS-Regeln wurde für jede Regel eine KS-Stelle definiert. Die KS-Stelle ist der Teil des Ausgangssatzes,

[17] https://www.tekom.de

[18] Die Tekom-Regeln sind der Kernregelsatz in zwei marktführenden CL-Checkern, nämlich Acrolinx (https://www.acrolinx.de/produktuberblick) und CLAT (vgl. Geldbach 2009). Mehr zum CLAT und seiner Entwicklung unter §2.6.

der bei dem Einsatz der KS-Regel modifiziert werden muss, und sein Äquivalent im Zielsatz. Beispielsweise wurde die Regel „Kurze Sätze formulieren“ ausgeschlossen, da sie sich auf den ganzen Satz bezieht. Entsprechend kann dafür keine konkrete KS-Stelle definiert werden, die bei der Fehlerannotation und der Humanevaluation analysiert und verglichen werden könnte (die KS-Stelle ist unten in diesem Abschnitt ausführlich erklärt).

Die Auswahl der Regeln aus der tekom-Leitlinie erfolgte in zwei Schritten: Im ersten Schritt [A] wurden die Regeln aus zwei definierten Gruppierungen der tekom, „Regeln für übersetzungsgerechtes Schreiben“ und „Basisregeln“, nach den obengenannten Kriterien ausgewählt. Da aus den beiden Gruppierungen der tekom nur vier KS-Regeln die Auswahlkriterien erfüllten, wurden alle KS-Regeln der tekom-Leitlinie im zweiten Schritt [B] geprüft, um weitere Regeln zu identifizieren, die die Auswahlkriterien erfüllen. Es folgt eine detaillierte Darstellung der beiden Schritte:

[A] In der aktuellen Auflage der Leitlinien von tekom findet man zwei für die Studie relevante Gruppierungen von Regeln:

Erste Gruppierung: Regeln für übersetzungsgerechtes Schreiben

Diese Gruppierung beinhaltet die Regeln, die u. a. für eine korrekte Verarbeitung durch Übersetzungswerkzeuge besonders relevant sind (tekom 2013: 136f.). In der folgenden Tabelle sind alle Regeln dieser Gruppierung enthalten (linke Spalte). In den anderen Spalten wird – zusammen mit der Begründung – erwähnt, welche Regeln dieser Gruppe analysiert bzw. ausgeschlossen wurden:

Übersetzungsgerechtes Schreiben	**Wurde die Regel analysiert? Wenn nicht, warum?**	
S 102 eindeutige pronominale Bezüge verwenden	✓	S 102 eindeutige pronominale Bezüge verwenden
S 204 keine Wortteile weglassen	✓	S 204 keine Wortteile weglassen
S 101 pronominale Bezüge über Satzgrenzen vermeiden	✗	Regel bezieht sich auf mehrere Sätze (erfüllt das 1. Auswahlkriterium nicht).
S 202 keine Sätze ohne Verb formulieren	✗	Dieser Verstoß kam im Korpus der Bedienungsanleitungen selten vor. Bsp.: Kühler defekt?

S 306 Aufzählungen als Listen darstellen S 307 Satz nicht durch eine Liste unterbrechen	✗	Die Regel bezieht sich in manchen Fällen zwar auf einen Satz (d. h. erfüllt das 1. Auswahlkriterium), jedoch geht der Satz über mehrere Zeilen. Dies würde die Humanevaluation erschweren, da die Probanden leicht von Stellen außerhalb der KS-Stelle abgelenkt werden können.
S 309 Klammereinschübe vermeiden	✗	Mehrere Umformulierungen sind möglich; meist durch die Bildung von zwei Sätzen (erfüllt das 2. Auswahlkriterium nicht).
Z 114 Verwendung von umbruchgeschützten Leerzeichen festlegen	✗	Dieser Verstoß kam im Korpus der Bedienungsanleitungen sehr selten vor.

Zweite Gruppierung: Basisregeln

Diese Gruppe bildet das Ergebnis zweier Umfragen ab, mit dem Ziel der Ermittlung der Regeln, die unter den tekom-Experten die größte Akzeptanz finden (tekom 2013: 144). In der folgenden Tabelle sind alle Regeln dieser Gruppierung enthalten (linke Spalte). In den anderen Spalten wird – zusammen mit der Begründung – erwähnt, welche Regeln dieser Gruppe analysiert bzw. ausgeschlossen wurden:

Basisregeln	**Wurde die Regel analysiert?**	
S 102 eindeutige pronominale Bezüge verwenden	✓	S 102 eindeutige pronominale Bezüge
S 201 Bedingungen mit “Wenn” formulieren	✓	S 201 Bedingungen mit “Wenn” formulieren
S 504: Passiv in bestimmten Informationseinheiten vermeiden	✓	S 504: Passiv in bestimmten Informationseinheiten vermeiden (Anweisungen, Sicherheits- und Warnhinweise)
T 101: einheitliche Überschriften definieren	✗	Regel bezieht sich auf ein komplettes Dokument (erfüllt das 1. Auswahlkriterium nicht).

S 103: missverständliche Genetivkonstruktionen vermeiden	✗	Einsatz der Regel kann auf verschiedene Weise erfolgen (erfüllt das 2. Auswahlkriterium nicht).
S 301: Häufung von Nominalphrasen vermeiden	✗	Einsatz der Regel kann auf verschiedene Weise erfolgen (erfüllt das 2. Auswahlkriterium nicht).
S 302: zu lange Sätze vermeiden	✗	Das Kürzen kann auf verschiedene Weise erfolgen (erfüllt das 2. Auswahlkriterium nicht).
S 304: Häufung von Präpositionalphrasen vermeiden	✗	Einsatz der Regel kann auf verschiedene Weise erfolgen (erfüllt das 2. Auswahlkriterium nicht).
S306: Aufzählungen als Listen darstellen S307: Satz nicht durch eine Liste unterbrechen	✗	Diese Regeln können sich zwar auf einen Satz beziehen (1. Auswahlkriterium), jedoch geht der Satz über mehrere Zeilen. Dies würde die Humanevaluation erschweren, da die Probanden leicht von Stellen außerhalb der KS-Stelle abgelenkt werden können.
S 311: Häufung von Nebensätzen vermeiden	✗	Einsatz der Regel kann auf verschiedene Weise erfolgen (erfüllt das 2. Auswahlkriterium nicht).
S 401: Sachlogische Reihenfolge einhalten	✗	Einsatz der Regel kann auf verschiedene Weise erfolgen (erfüllt das 2. Auswahlkriterium nicht).
S 505: Nominalisierungen vermeiden	✗	Satzstruktur muss – je nach Satz – angepasst werden; KS-Stelle ist schwer abzugrenzen (erfüllt das 3. Auswahlkriterium nicht).
S 510: einheitliche Satzmuster	✗	Regel bezieht sich auf mehrere Sätze (erfüllt das 1. Auswahlkriterium nicht).
B 101: Komposita aus zwei Basismorphem immer ohne Bindestrich	✗	Relevante Fälle waren im Korpus der Bedienungsanleitungen selten vertreten.

[B] Aus den obengenannten Gruppierungen haben vier KS-Regeln die Auswahlkriterien erfüllt. Daher wurden alle KS-Regeln der tekom-Leitlinie genauer geprüft, um weitere Regeln für die Analyse zu identifizieren. In der tekom-Leitlinie sind die Regeln wie folgt gegliedert:

Regeln	Wurde die Regel analysiert?	
Regeln zur Dokumentstruktur: Standardgliederung, einheitliche und kurze Überschriften, Indexeinträge, usw. (ebd.: 31 ff.)	✗	Regeln beziehen sich auf das gesamte Dokument (erfüllen das 1. Auswahlkriterium nicht).
Regeln zur Informationsstruktur i. S. v. Darstellung der Informationen (ebd.: 51 ff.)	✗	Allgemeine Regeln zur Darstellung von den Informationen in Tabellen, Glossaren, Listen, mit Aufzählungen usw. (erfüllen das 1. Auswahlkriterium nicht).
Satzregeln Diese Regeln stellen die Hauptzielgruppe der Analyse dar. (ebd.: 59 ff.)	✓	**(1) Regeln zur Vermeidung von mehrdeutigen Konstruktionen** S 102 eindeutige pronominale Bezüge verwenden **(2) Regeln zur Vermeidung von unvollständigen Konstruktionen** S 201 Bedingungen als „Wenn" –Sätze formulieren S 204 keine Wortteile weglassen **(3) Regeln zur Vermeidung von komplexen Konstruktionen** S 303 Partizipialkonstruktionen vermeiden **(4) Stilistische Regeln** S 501: Vorgangspassiv vermeiden S 502: Passiv mit Täterangabe vermeiden S 503: Passiv mit Modalverben vermeiden S 504 Passiv in bestimmten Informationseinheiten vermeiden S 511 Konstruktionen mit „sein + zu + Infinitiv" vermeiden

Satzregeln Diese Regeln stellen die Hauptzielgruppe der Analyse dar. (ebd.: 59 ff.)	✗	Die weiteren 32 Satzregeln erfüllen ein oder mehrere Auswahlkriterien nicht.
Wortregeln Diese Kategorie umfasst Regeln zur Wortbildung, Abkürzungen sowie Verwendung von Benennungen und Zahlen. (ebd.: 89 ff.)	✗	Die Analyse der Studie erfolgt auf Satzebene; die Wortebene ist kein Bestandteil der Studie.
Regeln zu den lexikalischen Vorgaben (ebd.: 106 ff.)	✓	L 103 Funktionsverbgefüge vermeiden L 114 Überflüssige Präfixe vermeiden
Regeln zu den lexikalischen Vorgaben (ebd.: 106 ff.)	✗	L 101 Keine ungenauen Verben verwenden Ungenaue Verben kamen im Korpus der Bedienungsanleitungen selten vor. Bsp.: machen, holen, geschehen
Regeln zur Rechtschreibung (ebd.: 114)	✗	R 101 Einheitlichen Rechtschreibstil verwenden (erfüllt das 2. Auswahlkriterium nicht).
Regeln zur Zeichensetzung (ebd.: 115 ff.)	✓	Z 103b: Für zitierte Oberflächentexte gerade Anführungszeichen "…" verwenden
Regeln zur Zeichensetzung (ebd.: 115 ff.)	✗	Die weiteren Regeln erfüllen ein oder mehrere Auswahlkriterien nicht bzw. relevante Fälle waren im Korpus der Bedienungsanleitungen selten vertreten.

Regeln zum Platzsparen beim Schreiben (ebd.: 129 ff.)	✗	Regeln wie P 101 Kurz formulieren, P 102 Kurze Wörter verwenden, P 104 Konsistenz halten usw. erfüllen ein oder mehrere Auswahlkriterien nicht bzw. beziehen sich auf Ebenen, die kein Bestandteil der vorliegenden Studie sind (z. B. Wortebene). Die Analyse der Studie erfolgt auf Satzebene.

Die KS-Stelle – Nachdem die zu analysierenden Regeln ausgewählt wurden, war es erforderlich, zu beobachten, auf welchen Teil der MÜ sie sich auswirken. Durch den Einsatz der einzelnen Regeln verändert sich der MÜ-Output sowohl semantisch als auch syntaktisch. Zu dieser Veränderung tragen mehrere Faktoren gleichzeitig bei, darunter der Ansatz des jeweiligen MÜ-Systems, die Trainingsdaten zusammen mit der Einführung der Kontrollierten Sprache. Da der Fokus der Studie auf dem Einfluss der KS-Regeln auf den MÜ-Output liegt, war es notwendig, die Stelle in der MÜ zu definieren, die direkt durch den Einsatz der KS-Regeln beeinflusst wird. Diese Stelle wird in der Studie als die *KS-Stelle* bezeichnet und wurde folgendermaßen definiert:

> Die KS-Stelle ist der Teil des Ausgangssatzes, der bei dem Einsatz der KS-Regel modifiziert werden muss, und sein Äquivalent im Zielsatz.

Im Zielsatz ist die KS-Stelle die exakte Übersetzung der Stelle, die durch den Einsatz der KS-Regel im Ausgangssatz umformuliert wurde. In anderen Worten ist sie die *kürzeste* Stelle im MÜ-Output, die durch den Einsatz der KS-Regel beeinflusst werden muss. Sollten weiteren Stellen im MÜ-Output vom KS-Einsatz beeinflusst werden, kämen vier Möglichkeiten in Betracht:

- Fall 1: Die KS-Stelle wurde *positiv* beeinflusst und außerhalb der KS-Stelle ebenfalls *positiv* beeinflusst → Ergebnis: Die KS-Regel hat einen starken *positiven* Einfluss;
- Fall 2: Die KS-Stelle wurde *negativ* beeinflusst und außerhalb der KS-Stelle *positiv* beeinflusst → Ergebnis: die KS-Regel hat einen *negativen* Einfluss;
- Fall 3: Die KS-Stelle wurde *negativ* beeinflusst und außerhalb der KS-Stelle ebenfalls *negativ* beeinflusst → Ergebnis: Die KS-Regel hat einen eindeutigen *negativen* Einfluss;

- oder Fall 4: Die KS-Stelle wurde *positiv* beeinflusst und außerhalb der KS-Stelle *negativ* beeinflusst → Ergebnis *muss geklärt werden.*

Auf Basis dieser Möglichkeiten liefert eine Analyse der KS-Stelle in den ersten drei Fällen eine ausreichende Aussage zum Einfluss der einzelnen Regeln. Der vierte Fall wurde bei den einzelnen Regeln (d. h. auf Regelebene) und den einzelnen Systemen (d. h. auf MÜ-Systemebene) näher untersucht (siehe §5.3.3).

Nachdem die Sätze maschinell übersetzt wurden, wurde das Äquivalent der KS-Stelle bei jeder Regel im Zielsatz identifiziert. Die Analyse der KS-Stelle – anhand der drei angewandten Methoden – ermöglichte es, den Einfluss der Regeln *einzeln* zu untersuchen. Im Vergleich zu Studien, in denen der allgemeine Einfluss der Kontrollierten Sprache auf den MÜ-Output untersucht wird, stellt die Untersuchung auf Regelebene eine besondere Schwierigkeit dar, denn bei einer Untersuchung des allgemeinen Einflusses der KS bewertet der Forscher den kompletten Output.

Die Idee der KS-Stelle ist ähnlich wie die der „Rich Points", die die PACTE[19] Group in Anlehnung an das Konzept der „Wissenschaftlichen Ökonomie" (Scientific Economy) von Giegler anwendet (Beeby u. a. 2011). Bei den Rich Points handelt es sich um „specific source-text segments that contained ‚prototypical' translation problems" (ebd.: 10). Zur Erleichterung der Datenerfassung und -analyse im Bereich der Translationskompetenz und -evaluation legt die PACTE Group bei ihrer Untersuchung zur Identifizierung und Lösung von Translationsproblemen (ebd.) den Fokus auf die „Rich Points".

4.5.2.2 Darstellung der analysierten KS-Regeln und ihrer gezielten Wirkung

Dieser Abschnitt bietet eine detaillierte Darstellung der analysierten Regeln: Zuerst werden die analysierten tekom-Regeln beschrieben. Außerdem wird die Begründung der Regelanwendung laut der tekom sowie die gezielte Wirkung jeder Regel laut vorherigen Studien angegeben. Danach wird demonstriert, wie die Regel umgesetzt wurde (Umsetzungsmuster). Anschließend wird die KS-Stelle vor und nach dem KS-Einsatz spezifiziert und anhand eines Beispiels vorgestellt.

Z 103b: Für zitierte Oberflächentexte gerade Anführungszeichen "..." verwenden Nach dieser Regel sollen Oberflächentexte, z. B. Texte in Softwareoberflächen oder Displaytexte in Geräten, in geraden Anführungszeichen stehen (tekom 2013: 117).

[19] PACTE steht für Process in the Acquisition of Translation Competence and Evaluation (Beeby u. a. 2011).

Begründung der Anwendung laut tekom: Die Anführungszeichen erhöhen die Lesbarkeit. Im Vergleich zu der Verwendung von verschiedenen Schriftarten oder Schriftgraden sind gerade Anführungszeichen optisch nicht störend. Zudem unterstützen die Anführungszeichen eine korrekte Übersetzung. (ebd.: 118)

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:
Eine Wirkung auf die MÜ ist nachvollziehbar, denn laut Reuther (2003: 2): „Punctuation marks are very sensitive with respect to all applications where linguistic processing is done automatically."

Umsetzungsmuster:
Vor-KS: Oberflächentext ohne Anführungszeichen
Nach-KS: Oberflächentext angegeben in geraden Anführungszeichen

KS-Stelle:
Vor-KS: Oberflächentext ohne Anführungszeichen
Nach-KS: Oberflächentext mit geraden Anführungszeichen

Beispiele:
Wählen Sie danach die Option *Software automatisch installieren* .
Wählen Sie danach die Option *"Software automatisch installieren"* .

L 103: Funktionsverbgefüge vermeiden Nach dieser Regel soll das bedeutungstragende Verb anstatt des Funktionsverbgefüges verwendet werden (tekom 2013: 107).

Begründung der Anwendung laut tekom: Die Verwendung des bedeutungstragenden Verbs steigert die Präzision und Direktheit der Aussage (ebd.).

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:
Das Vermeiden ausdrucksschwacher Verben reduziert die Ambiguität (Siegel 2011). Formulierungen wie das Funktionsverbgefüge „machen den Sachverhalt unnötig kompliziert und erschweren das Textverständnis und die Übersetzung." (Congree 2018: 13)

Umsetzungsmuster:

Vor-KS: Funktionsverbgefüge
Nach-KS: Das Funktionsverbgefüge wird durch das bedeutungstragende Verb ersetzt.

KS-Stelle:

Vor-KS: Funktionsverbgefüge
Nach-KS: bedeutungstragendes Verb

Beispiele:

Im oberen Abschnitt können Sie Einstellungen für die angezeigten Module vornehmen.

Im oberen Abschnitt können Sie die angezeigten Module einstellen.

S 201: Konditionalsätze mit ‚Wenn' einleiten Nach dieser Regel (tekom 2013: 66; auch bekannt als „Bedingungen als ‚Wenn'-Sätze formulieren") sollen Konditionalsätze mit der Konjunktion ‚wenn' oder ‚falls' eingeleitet werden.

Begründung der Anwendung laut tekom: Durch die Satzstruktur „Wenn-Nebensatz-Hauptsatz" wird das „Bedingung-Folge-Verhältnis" ersichtlich und somit die Textverständlichkeit erhöht (ebd.: 67).

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:

Die Regel sollte das Parsing aufgrund der Komplexität der Übersetzung elliptischer Konstruktionen verbessern. Reuther (2003: 3) thematisiert die Komplexität der Übersetzung elliptischer Konstruktionen wie folgt: „Human and machine parsing mechanisms have to reconstruct the missing elements, which results in readability problems or, in the case of MT systems, in failed parses."

Umsetzungsmuster:

Vor-KS: Der Satz beginnt mit dem Verb.
Nach-KS: Der Satz ist mit ‚Wenn' eingeleitet. Wenn der Satz vor-KS mit ‚so' formuliert ist, wurde ‚so' aus stilistischen Gründen nach-KS entfernt.

KS-Stelle:

Vor-KS: Verb am Satzanfang
Nach-KS: ‚wenn' + Verb

Beispiele:

Ist die Seriennummer des Gerätes bekannt, kann im Feld ...

Wenn die Seriennummer des Gerätes bekannt ist, kann im Feld ...

Werden die vordefinierten Werte verändert, so erfolgt die Umrechnung automatisch.

Wenn die vordefinierten Werte verändert werden, erfolgt die Umrechnung automatisch.

S102: Eindeutige pronominale Bezüge verwenden Nach dieser Regel (tekom 2013: 60) sollen Pronomen vermieden werden, wenn sie mehrdeutig sein könnten. Anstelle des Pronomens soll das Bezugswort wiederholt werden, damit der Bezug eindeutig erkennbar wird.

Begründung der Anwendung laut tekom: Der Leser eines technischen Dokuments ist in der Regel weniger mit der Thematik vertraut als der Autor (ebd.: 61). Ferner wird diese Regel von der tekom (ebd.: 137) für ein übersetzungsgerechtes Schreiben empfohlen.

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:

Diese Regel wird von Congree (2018) zur Vermeidung von Mehrdeutigkeit empfohlen. Gleichzeitig spielen Pronomen eine wesentliche Rolle in der natürlichen Sprache, sodass der Autor bei der Anwendung der Regel fallabhängig zwischen der maschinellen Übersetzbarkeit und einem natürlich klingenden Text abwägen muss (Bernth & Gdaniec 2001).

Umsetzungsmuster:

Vor-KS: Der Satz beinhaltet ein Pronomen.
Nach-KS: Das Pronomen wird durch das Nomen ersetzt.

KS-Stelle:

Vor-KS: Pronomen (Personalpronomen und Demonstrativpronomen)
Nach-KS: Nomen bzw. Demonstrativpronomen und Nomen (inkl. damit verbundener Fehler in der Wortstellung)[20]

Beispiele:

Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit,

ihn rückstandslos zu entfernen.
Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, den Fleck rückstandslos zu entfernen.

Sofern auf der Oberfläche alte Kleberreste anhaften, sind diese vollständig zu entfernen.
Sofern auf der Oberfläche alte Kleberreste anhaften, sind diese Kleberreste vollständig zu entfernen.

S 303: Partizipialkonstruktion vermeiden Nach dieser Regel (tekom 2013: 70) soll statt der Partizipialkonstruktion eine einfache Satzstruktur mit mehreren kurzen Sätzen oder Nebensätzen verwendet werden.

Begründung der Anwendung laut tekom: Die Partizipialkonstruktion erschwert die Textverständlichkeit (ebd.).

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:
Komplexe Satzstrukturen, wie die Partizipialkonstruktionen, sind sowohl für den Menschen als auch das Maschinen-Parsing problematisch, sodass sie die Lesbarkeit und die Übersetzbarkeit im Kontext der MÜ beeinträchtigen (Reuther 2003). Daher wurde diese Regel zur Förderung der maschinellen Übersetzbarkeit von Bernth & Gdaniec (2001) aufgeführt.

Umsetzungsmuster:
Vor-KS: Der Satz beinhaltet eine Partizipialkonstruktion.
Nach-KS: Die Partizipialkonstruktion wird in einem Nebensatz aufgelöst.

KS-Stelle:
Vor-KS: alle Wörter, die das Nomen beschreiben, angefangen mit dem Artikel, falls vorhanden
Nach-KS: die konvertierten Wörter von der Version vor-KS (inkl. des Kommas innerhalb der KS-Stelle)

Beispiele:
Speziell auf diese Lautsprecher abgestimmtes Zubehör erhalten Sie in unserem Webshop.

Zubehör, das speziell auf diese Lautsprecher abgestimmt ist , erhalten Sie in unserem Webshop.

Passiv vermeiden

S 501: Vorgangspassiv vermeiden
S 502: Passiv mit Täterangabe vermeiden
S 503: Passiv mit Modalverben vermeiden
S 504: Passiv in bestimmten Informationseinheiten vermeiden
Nach diesen vier Regeln (tekom 2013: 79 ff.) soll die Verwendung des Passivs vermieden und stattdessen sollen die Sätze in Aktiv formuliert werden.

Begründung der Anwendung laut tekom: Die Passivkonstruktion ist oft nicht eindeutig und lässt den Handelnden unklar. Wenn der Täter genannt werden soll, eignet sich die Aktivformulierung. (ebd.: 79) Insbesondere bei Anweisungen, Sicherheits- und Warnhinweisen wird die Aktivkonstruktion empfohlen, damit dem Leser klar wird, wer die Handlung ausführt oder ausführen soll. Warnungen wirken motivierend, wenn sie im Aktiv formuliert sind, da der Leser direkt angesprochen wird. (ebd.: 81)

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:
Das Vermeiden des Passivs führt zur Verbesserung der MÜ-Qualität (Siegel 2013).[21] Reuther (2003) empfahl diese Regel, um Parsing-Probleme umgehen zu können. Bernth & Gdaniec (2001: 190) erkennen: „Passive voice plays a role in creating the right focus in a sentence, among other things." Sie empfehlen gleichzeitig, das Passiv zum Zweck einer besseren maschinellen Übersetzbarkeit zu vermeiden, wenn es aus stilistischer Sicht nicht notwendig sei (ebd.).

Umsetzungsmuster:
Vor-KS: Satz formuliert in Passiv
Nach-KS: Satz formuliert in Aktiv

KS-Stelle:
Vor-KS: Form von ‚werden' + Partizip II
Nach-KS: Verb bzw. Subjekt + Verb, wenn das Subjekt im Passivsatz nicht enthalten war und erst im Aktivsatz hinzugefügt wurde

Beispiele:

Bei der Arbeit mit elektrischen Geräten sollte stets ein Sicherheitsstecker verwendet werden.

Bei der Arbeit mit elektrischen Geräten verwenden Sie stets einen Sicherheitsstecker.

Das Programm wird vom Hersteller wie folgt eingestellt.

Der Hersteller stellt das Programm wie folgt ein.

S 511: Konstruktionen mit „sein + zu + Infinitiv" vermeiden Nach dieser Regel (tekom 2013: 86) soll die Passiv-Ersatzkonstruktion „sein + zu + Infinitiv" bei Anweisungen vermieden werden. Stattdessen sollen sie durch einen Infinitiv oder direkte Anrede formuliert werden.

Begründung der Anwendung laut tekom: Diese Passiversatzkonstruktion ist umständlich und der Leser wird nicht direkt angesprochen. Der Infinitiv bzw. eine direkte Anrede fördert die schnelle und richtige Umsetzung der Handlung (ebd.).

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:

Diese Regel bzw. eine Formulierung mithilfe des Imperativs wird von Congree (2018) für eine präzise und deutliche Anweisung empfohlen.

Umsetzungsmuster:

Vor-KS: Der Satz ist mit einem Passiversatz (sein + zu und das Verb am Satzende) formuliert.
Nach-KS: zwei mögliche Varianten:

- Imperativ am Satzende
- Imperativ am Satzanfang

Beide Varianten wurden bei der Übersetzung mit MÜ-Systemen getestet. Da die erste Variante mit mehr Fehlern in dem MÜ-Output verbunden war, wurde entschieden, die zweite Variante als Einsatzmuster zu verwenden. Wenn der Satz vor-KS mit ‚so' formuliert war, wurde ‚so' aus stilistischen Gründen nach-KS entfernt.

KS-Stelle:

Vor-KS: sein + zu + das Verb am Satzende
Nach-KS: Imperativ + Subjekt

Beispiele:

Die Herstellerangaben sind stets zu beachten .

Beachten Sie stets die Herstellerangaben.

Ist ein mehrstufiges Modul parametriert, so sind die externen Kontakte zu verriegeln .

Ist ein mehrstufiges Modul parametriert, verriegeln Sie die externen Kontakte.

L 114: Überflüssige Präfixe vermeiden Nach dieser Regel sollen Verben mit Präfixen vermieden werden, wenn das Verb ohne Präfix die gleiche Bedeutung hat (tekom 2013: 111).

Begründung der Anwendung laut tekom: Die Kürzung vereinfacht den Satz und reduziert Segmentvarianten bei der MÜ (ebd.).

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:

Die Verwendung trennbarer Verben in ihrer getrennten Form erhöht die Komplexität der Satzstruktur (Siegel 2011). Diese Regel führt zur Reduzierung der Ambiguität (Bernth & Gdaniec 2001) und damit zur Verbesserung der maschinellen Übersetzbarkeit (ebd.) bzw. Erhöhung der MÜ-Qualität (Siegel 2013).

Umsetzungsmuster:

Vor-KS: Verb mit Präfix
Nach-KS: Eliminierung des Präfixes

KS-Stelle:

Vor-KS: Verb mit Präfix (trennbare und untrennbare Verben)
Nach-KS: Verb ohne Präfix

Beispiele:

Überprüfen *Sie, ob ausreichend Wasser im Wassertank vorhanden ist.*

Prüfen *Sie, ob ausreichend Wasser im Wassertank vorhanden ist.*

Wählen *Sie die Option "Software von einer bestimmten Liste installieren"* *aus* *.*

Wählen *Sie die Option "Software von einer bestimmten Liste installieren".*

S 204: Keine Wortteile weglassen Nach dieser Regel sollen Wörter vollständig ausgeschrieben werden (tekom 2013: 68).

Begründung der Anwendung laut tekom: Ziel der Regel ist die Unterstützung des Verständnisses. Insbesondere bei der Übersetzung ist das Weglassen von Wortteilen ungeeignet. (ebd.)

Begründung der Anwendung bzw. die gezielte Wirkung der Regel laut vorherigen Studien:

Durch die Beschreibung der Ellipsen von Halliday und Hasan (Bernth & Gdaniec 2001: 189) als „specific structural slots to be filled by elsewhere" wird die Problematik ihrer MÜ ersichtlich. Die MÜ-Systeme verfügen nicht immer über Quellen zum Füllen der fehlenden Slots, daher führen die tekom (2013: 68) sowie mehrere weitere Regelsätze (Bernth & Gdaniec 2001; Reuther 2003; Siegel 2011; Congree 2018) diese Regel zur Unterstützung der maschinellen Übersetzbarkeit, Reduzierung der Ambiguität sowie ggf. Erhöhung der Textverständlichkeit auf.

Umsetzungsmuster:

Vor-KS: Der Satz beinhaltet Wortteile.
Nach-KS: Die fehlenden Wortteile werden vervollständigt.

KS-Stelle:

Vor-KS: Wörter mit weggelassenen Teilen
Nach-KS: vollständige Wörter

Beispiele:

Die *Ist- und Sollwerte* *des zweiten Regelkreises werden nach der Konfiguration angezeigt.*

> *Der Istwert und der Sollwert des zweiten Regelkreises werden nach der Konfiguration angezeigt.*

Wie die obige Darstellung zeigt, haben nicht alle analysierten Regeln nach der tekom (2013) die Übersetzbarkeit im Fokus; einige Regeln zielen auf die Lesbarkeit bzw. die Verständlichkeit ab. Jedoch zeigen die weiteren zitierten Studien (Bernth & Gdaniec 2001; Reuther 2003; Siegel 2011; Congree 2018), dass die analysierten Regeln durch die Reduzierung der Satzkomplexität, Vereinfachung der Satzstruktur bzw. Verminderung der Ambiguität die maschinelle Übersetzbarkeit – bei den früheren MÜ-Ansätzen - verbessern. Auf diesem Wege tragen auch Regeln, die die Verständlichkeit in den Mittelpunkt stellen, indirekt zur Verbesserung der Übersetzbarkeit bei. Diese allgemeine Wirkung der KS-Regeln auf die MÜ wurde von Fiederer & O'Brien (2009: 53) wie folgt bestätigt:

> Another method for improving MT output is to apply Controlled Language (CL) rules to the source text in order to reduce ambiguities and complexity. CL rules generally make the source text input more suitable for MT by reducing sentence length and eliminating problematic features. (Fiederer & O'Brien 2009: 53)

In ihrer Studie mit dem Titel „Two in one – Can it work?" bildete Reuther (2003) zwei KS-Regelsätze, den ersten Regelsatz zur Verbesserung der Lesbarkeit und Verständlichkeit und den zweiten zur Verbesserung der Übersetzbarkeit. Sie fand heraus, dass sich die beiden Regelsätze nicht allzu sehr unterschieden. Die Regeln der Lesbarkeit und Verständlichkeit waren eine Teilmenge der Regeln der Übersetzbarkeit. Auf Basis dieser Ergebnisse untersuchte Reuther, wie eine KS-Konformitätsprüfung (CL-Check) in der Praxis am effizientesten durchgeführt werden kann; sie empfahl eine gemeinsame automatisierte Verarbeitung (common automated processing) der Verständlichkeitsprüfung und der Übersetzbarkeitsprüfung. (ebd.) Vor diesem Hintergrund wird der Einfluss der aufgeführten Regeln auf die MÜ-Qualität in der vorliegenden Studie empirisch untersucht.

In dieser empirischen Untersuchung wird der Output von vier MÜ-Ansätzen, darunter des NMÜ-Ansatzes, verglichen. Wie bisherige Studien zeigen, verzeichnet der NMÜ-Ansatz einen großen Fortschritt bei der Erstellung von Output,

[20] Siehe Berechnungsvorgehensweise von Wortstellungsfehlern in §4.5.4.1 unter „Die Fehlertypen".

[21] Die MÜ-Qualität wurde auf einer Skala von nn bis 3, in der nn für „Verbesserung nicht erforderlich"; 1 für „keine Verbesserung"; 2 für „leichte Verbesserung"; 3 für „starke Verbesserung" stehen, gemessen.

der deutlich flüssiger ist und weniger Morphologie- und Grammatikfehler im Vergleich zum davor dominanten Ansatz, PBMÜ, aufweist (vgl. Bentivogli u. a. 2016; Klubička u. a. 2017; Toral & Sanchez-Cartagena 2017). Gleichzeitig wird an dem NMÜ-Output kritisiert, dass seine hohe Flüssigkeit in manchen Fällen mit einer bedenklichen Adäquatheit einhergehe (vgl. Castilho, Moorkens, Gaspari, Calixto u. a. 2017; Koehn 2017). Dementsprechend erfordere das Post-Editing des NMÜ-Outputs eine hohe Präzision bzw. einen hohen kognitiven Aufwand, um Adäquatheits- und stilistische Fehler zu identifizieren und zu korrigieren (vgl. Volk 2018; Vardaro u. a. 2019). Angesichts dieser Entwicklung dürfen die Adäquatheit und der Stil bei einer MÜ-Evaluation nicht außer Acht gelassen werden. Daher nimmt die vorliegende Studie die einzelnen Regeln unter die Lupe, um ihre Wirkung auf die MÜ-Qualität sowohl stilistisch als auch inhaltlich im Sinne der Verständlichkeit und Genauigkeit empirisch zu testen.

4.5.2.3 Diskussion der analysierten Regeln und ihrer gezielten Wirkung im Hinblick auf das untersuchte Sprachenpaar

Nachdem die analysierten Regeln dargestellt wurden (§4.5.2.2), werden sie und ihre gezielten Wirkungen in diesem Abschnitt im Hinblick auf das untersuchte Sprachenpaar (Deutsch-Englisch) diskutiert. Die Studienergebnisse hängen weitgehend vom untersuchten Sprachenpaar ab. Ein Vergleich der MÜ-Outputs von Sprachenpaaren wie Deutsch <> Arabisch, Englisch <> Arabisch und Deutsch <> Englisch zeigt, dass die MÜ-Systeme bei dem Sprachenpaar Deutsch <> Englisch einen deutlich besseren Output liefern. Dies ist auf mehrere Gründe zurückzuführen, darunter die Entwicklungs- und Forschungsintensität sowie die Verfügbarkeit von zweisprachigen Korpora bzw. Trainingsdaten. Diese Aspekte tragen zwar zu einem besseren Output im Falle des Sprachenpaars Deutsch-Englisch bei, dennoch bleiben die Sprachunterschiede ein Hindernis, dessen Bewältigungsgrad von vielen Faktoren wie der Domäne, dem Texttyp und der Textkomplexität abhängt.

Das analysierte Sprachenpaar Deutsch > Englisch weist mehrere bekannte Unterschiede auf. Die deutsche Sprache verfügt über eine reichere Flexionsmorphologie im Vergleich zur englischen Sprache (Hawkins 1986: 11). Sämtliche grammatischen Unterscheidungen innerhalb der englischen Flexionsmorphologie sind im Deutschen vertreten, umgekehrt ist dies aber nicht der Fall (ebd.). Diese Tatsache kann aber aus translatorischer Sicht vorteilhaft sein. Alleine die Unterscheidung zwischen vier Kasus im Deutschen (Nominativ, Akkusativ, Dativ und Genitiv) gegenüber nur einem einzigen Kasus im Englischen stellt eine große Vereinfachung bei der Übersetzung aus dem Deutschen ins Englische dar. Ein weiterer

Unterschied besteht darin, dass die deutsche Sprache im Gegensatz zur englischen Sprache über einen höheren Grad an Wortstellungsfreiheit – auch Grad der „configurationality“ genannt – verfügt (vgl. Hawkins 1986: 37ff.). Englisch wird als „configurational“ Sprache betrachtet, da es eine starre Wortstellung hat, während Deutsch mehr Freiheit bei der Wortstellung – im Rahmen seiner Satzregeln – anbietet und entsprechend als „less-configurational“ Sprache klassifiziert wird. Da die Studie sich mit der Übersetzung aus dem Deutschen ins Englische befasst, kann die eingeschränkte „configurationality“ der deutschen Sprache eine Übersetzungsschwierigkeit darstellen. Denn je starrer die Wortstellung eines Satzes ist, desto einfacher ist es für die MÜ-Systeme, ihn zu parsen.

Im Folgenden wird auf die sprachlichen Unterschiede bei den untersuchten Regeln eingegangen und entsprechend der Studienfokus in Bezug auf den Einfluss der einzelnen Regeln auf den MÜ-Output abgeleitet:

4.5.2.3.1 Regel „Für zitierte Oberflächentexte gerade Anführungszeichen "…" verwenden“

Laut Duden können Anführungszeichen im Deutschen zur Hervorhebung verwendet werden.[22] Orthografisch haben Anführungszeichen nach Nerius (2007: 254) die Rolle der Kennzeichnung fremder Äußerungen, wie direkter Rede und Zitaten, aber auch Titeln und Überschriften innerhalb des Satzverbandes. Oberflächentexte sind „alle Texte einer Softwareoberfläche oder Texte, die sich auf einem Gerät befinden“ (tekom 2013: 117) und gelten somit als fremde Äußerungen innerhalb des Satzverbandes. Auf dieser Basis wird diese Regel von der tekom zur Erhöhung der Lesbarkeit wegen der verbesserten Orientierung empfohlen. Dabei haben die Anführungszeichen gegenüber weiteren Hervorhebungstechniken (wie der Verwendung von anderen Schriftarten, Schriftgraden oder Schriftschnitten) laut tekom den Vorteil, dass sie das Schriftbild weniger beeinflussen. (tekom 2013: 118) Im Englischen hingegen ist die Verwendung von Anführungszeichen keine gängige Hervorhebungstechnik. McMurrey (2006) empfiehlt den technischen Redakteuren: „limit quotation marks to the traditional usage, which includes quoted speech; numbers, letters, or words referred to as such“; seine Begründung dafür lautet: „Quotation marks, like capital letters, tend to create a busy, distracting text and therefore should be avoided“ (ebd.).

Eine korrekte Anwendung dieser Regel fördert der tekom zufolge (2013: 118) „nicht nur die Konsistenz, sondern auch die Übersetzung.“ Dies ist nachvollziehbar, denn laut Reuther (2003: 2): „Punctuation marks are very sensitive with re-

[22] Online unter: https://www.duden.de/sprachwissen/rechtschreibregeln/anfuehrungszeichen#D8.

spect to all applications where linguistic processing is done automatically." Dementsprechend kann diese Regel die MÜ auf zweierlei Weise beeinflussen: Einerseits – anders als im Deutschen – kann das Auftreten der Anführungszeichen in der englischen MÜ aufgrund der untypischen Verwendung als optisch störend bzw. ablenkend wahrgenommen werden. Andererseits hat die Verwendung von Anführungszeichen in Zusammenhang mit der MÜ im Gegensatz zu einer Hervorhebung mittels der Verwendung anderer Schriftarten oder Formatierungen (z. B. fett, kursiv) den Vorteil, dass die Grenze des Oberflächentexts innerhalb des Satzverbandes gekennzeichnet wird (vgl. Nerius 2007: 253). Diese Kennzeichnung kann die MÜ-Systeme bei der Tokenisierung und dem Parsen unterstützen, was wiederum eine korrekte Übersetzung fördert, wie von der tekom (2013: 118) beabsichtigt. In der Studie wird entsprechend untersucht, inwiefern die Anwendung der Anführungszeichen bei der Tokenisierung und Erkennung der Oberflächentexte als feste bzw. spezifische Begriffe, die oft als Mehrwortentitäten auftreten, förderlich ist und wie sie sich auf die MÜ-Qualität auswirkt.

4.5.2.3.2 Regel „Funktionsverbgefüge vermeiden"

Generell wird das Funktionsverbgefüge kritisiert und u. a. als „Folterinstrument für Verben und Leser" bezeichnet (Baumert & Verhein-Jarren 2012: 106). Heine (2017) stellt Funktionsverbgefüge als „ein typisches Beispiel für Phänomene, die weder (ausschließlich) mit grammatischen Regeln noch als lexikalische Einheiten erklärbar sind" dar und verdeutlicht wie die Satzsyntax zusammen mit der lexikalischen Komponente des MÜ-Systems entscheidend für einen fehlerfreien MÜ-Output sind. Das Vermeiden des Funktionsverbgefüges bzw. ausdrucksschwacher Verben ist in mehreren KS-Regelsätzen enthalten. Die tekom (2013: 107) empfiehlt diese Regel, um die Präzision und Direktheit der Aussage zu steigern. Weitere Studien zeigen, dass das Vermeiden ausdrucksschwacher Verben die Ambiguität reduziert (Siegel 2011). Außerdem ist diese Regel ein Bestandteil der Stilregeln von Congree (2018) und wird mit der folgenden Begründung empfohlen: „Solche Formulierungen machen den Sachverhalt unnötig kompliziert und erschweren das Textverständnis und die Übersetzung." (ebd.: 13)[23]

Das Funktionsverbgefüge im Deutschen ist mit der „light verb construction" im Englischen vergleichbar. Trotz des Vorhandenseins in beiden Sprachen wird seine Übersetzung aufgrund der Transferprobleme als Herausforderung betrachtet (vgl. Baumert & Verhein-Jarren 2012: 107). Einige Funktionsverbgefüge haben

[23] Congree Language Technologies GmbH ist eines der marktführenden Unternehmen für KS-Softwareprodukte.

Pendants im Englischen (z. B. ‚zur Entscheidung kommen' → ‚to make a decision'); bei anderen Funktionsverbgefügen ist dies nicht der Fall (z. B. ‚in Verbindung setzen'). Hat das Funktionsverbgefüge ein Pendant im Englischen und ist die Satzstruktur *nicht* komplex, so ist eine korrekte MÜ des Funktionsverbgefüges zu erwarten. Hat das Funktionsverbgefüge ein Pendant im Englischen und ist die Satzstruktur komplex, kann dies in einer fehlerhaften MÜ resultieren. In dem Fall würde die Regelanwendung durch die Verwendung des bedeutungstragenden Verbs die Satzstruktur vereinfachen, was zur Erleichterung des Parsens und wiederum zur Verbesserung des MÜ-Outputs beitragen würde. Hat das Funktionsverbgefüge kein Pendant im Englischen, dann sollte ebenfalls die Anwendung der Regel durch die Reduzierung der Ambiguität und die Vereinfachung der Satzstruktur (vgl. Siegel 2011; Congree 2018) hilfreich sein. Außerdem kann der Effekt, der von der tekom (2013: 107) bei der Ausgangssprache, beabsichtigt ist, nämlich den Satz konkreter und direkter zu gestalten, sich in der Zielsprache widerspiegeln. Dies wiederum würde dazu beitragen, dass die MÜ aufmerksamkeitserregend wirkt.

Trotz der damit einhergehenden Schwierigkeiten ist ein vollständiger Verzicht auf das Funktionsverbgefüge nicht möglich, da „manche Funktionsverbgefüge Bedeutungsnuancen tragen, die mit dem einfachen Verb schwer oder gar nicht gesetzt werden können" (z. B. kann ‚eine Maschine in Betrieb zu setzen' komplexe Verfahren beinhalten und ist somit in manchen Fällen nicht mit ‚einschalten' gleichzustellen) (Baumert & Verhein-Jarren 2012: 107). „Für andere Funktionsverbgefüge fehlt die Entsprechung völlig" (z. B. ‚in Ordnung halten') (ebd.). Somit begründet das potenzielle Fehlen eines äquivalenten einfachen Verbes die Notwendigkeit der Verwendung des Funktionsverbgefüges. Vor dem Hintergrund untersucht die Studie, inwiefern der bisher realisierte MÜ-Fortschritt eine korrekte Übersetzung des Funktionsverbgefüges ermöglicht.

4.5.2.3.3 Regel „Konditionalsätze mit ‚Wenn' einleiten"

Anders als im Deutschen, in dem der Konditionalsatz mit dem Verb eingeleitet werden kann, beginnen die Konditionalsätze im Englischen in der Regel mit ‚If'. Ausnahmen sind nur in der Umgangssprache oder den literarischen Inversionsstrukturen zu finden (Swan 1980: 307): Im Umgangsenglisch kann das ‚If' weggelassen werden, während in den literarischen Inversionsstrukturen der Konditionalsatz mit ‚Were', ‚Should' und ‚Had' anstelle von ‚If' beginnen kann. Neben den sprachlichen Unterschieden thematisiert Reuther (2003: 3) die Komplexität der Übersetzung elliptischer Konstruktionen wie folgt: „Human and machine parsing

mechanisms have to reconstruct the missing elements, which results in readability problems or, in the case of MT systems, in failed parses.“ Aufgrund der sprachlichen Unterschiede und der Parsing-Problematik von elliptischen Konstruktionen sollte die Regelanwendung die Satzstruktur eindeutiger gestalten und somit die Systeme dabei unterstützen, die Konditionalsätze korrekt ins Englische zu übersetzen. In der Studie wird untersucht, inwiefern die MÜ einen Fortschritt bei dieser Art der elliptischen Konstruktion gemacht hat bzw. ob die Regelanwendung für eine fehlerfreie Übersetzung erforderlich ist.

4.5.2.3.4 Regel „Eindeutige pronominale Bezüge verwenden“

Pronominale Bezüge lassen sich in beiden Sprachen wiederholen. Einige Pronomen im Deutschen, wie das Personalpronomen ‚sie‘ oder das Demonstrativpronomen ‚diese‘, können – je nach Satzstruktur – mehrdeutig sein, denn die Erkennung des Pronomenbezugs (sog. Koreferenzauflösung)[24] stellt in der MÜ eine Schwierigkeit dar (vgl. Ng 2017). Daher wird diese Regel von der tekom (2013: 137) für ein übersetzungsgerechtes Schreiben bzw. von Congree (2018) zur Vermeidung von Mehrdeutigkeit empfohlen. Gleichzeitig spielen Pronomen eine wesentliche Rolle in der natürlichen Sprache, sodass der Autor bei der Anwendung der Regel fallabhängig zwischen der maschinellen Übersetzbarkeit und einem natürlich klingenden Text abwägen muss (Bernth & Gdaniec 2001). Die Studie untersucht, ob diese Abwägung nach jetzigem Entwicklungsstand der MÜ weiterhin notwendig ist. In anderen Worten, ob sich in der Zwischenzeit die Koreferenzauflösung verbessert hat, wodurch die Begünstigung eines natürlichen Stils für gewünschte Textsorten möglich ist.

4.5.2.3.5 Regel „Partizipialkonstruktion vermeiden“

Sowohl im Deutschen als auch im Englischen wird die Partizipialkonstruktion in der Funktion eines Relativsatzes verwendet (vgl. Königs 2004: 182ff.). In beiden Sprachen wird diese Art der Partizipialkonstruktion zur Sprachökonomie verwendet; zudem wird sie im Deutschen als gehobener Stil empfunden (ebd.). Vor diesem Hintergrund ist es erforderlich, dass die MÜ-Systeme (zumindest einfache) Partizipialkonstruktionen korrekt übersetzen können. Dennoch zählen Partizipialkonstruktionen zu den komplexen Satzstrukturen. Komplexe Satzstrukturen sind sowohl für den Menschen als auch das Maschinen-Parsing problematisch, sodass sie die Lesbarkeit und die Übersetzbarkeit im Kontext der MÜ beein-

[24] Durch die Koreferenzauflösung wird die Entität, worauf sich die Koreferenz bzw. das Pronomen bezieht, identifiziert (vgl. Ng 2017).

trächtigen (Reuther 2003). Daher wurde diese Regel[25] zur Förderung der maschinellen Übersetzbarkeit von Bernth & Gdaniec (2001) aufgeführt. Die Regel kann dementsprechend insbesondere bei komplexen Partizipialkonstruktionen nützlich sein, um die Satzstruktur sowohl für den Menschen als auch das Maschinen-Parsing zu vereinfachen. In der Studie wird die Nützlichkeit der Regel nach dem derzeitigen Entwicklungsstand der MÜ untersucht und ihre Auswirkung auf die MÜ inhaltlich und stilistisch gemessen.

4.5.2.3.6 Regel „Passiv vermeiden“

„Kaum eine Debatte um sprachliche Fragen in der Technischen Dokumentation wird so hartnäckig und engagiert geführt wie die der Passiv-Verwendung.“ (Muthig 2003) Die Verwendung des Passivs ist ein häufig diskutiertes Thema in der technischen Redaktion sowohl im Englischen als auch im Deutschen. Die Regel „Passiv vermeiden“ ist ebenfalls ein Bestandteil der tekom-Leitlinie für die englische Sprache „Regelbasiertes Schreiben – Englisch für deutschsprachige Autoren“ (Siegel 2014). Im Aktiv wird der Leser direkt angesprochen und zum Handeln angeregt, daher empfiehlt die tekom (2013: 81) seine Anwendung bei der Formulierung von handlungsorientierten Informationseinheiten, wie z. B. Sicherheitshinweisen. Im Passiv hingegen steht die Handlung im Mittelpunkt. (vgl. Baumert & Verhein-Jarren 2012: 67f.) Die Verwendung vom Passiv kann sinnvoll sein, „wenn der Handelnde nicht bekannt ist oder bewusst nicht genannt werden soll“ (tekom 2013: 80). Auf dieser Basis sollten die beiden Konstruktionen in der technischen Dokumentation unterschiedlich je nach Kontext und Satzintention eingesetzt werden. Dementsprechend ist es notwendig, dass die MÜ-Systeme beide Formen – bei klarer Satzstruktur und adäquatem Terminologiemanagement – fehlerfrei übersetzen können.

Bisherige Studien fanden jedoch heraus, dass das Vermeiden des Passivs zu einer „starken Verbesserung“ der MÜ-Qualität führt (Siegel 2013).[26] Reuther (2003) empfahl diese Regel, um Parsing-Probleme umgehen zu können. Bernth & Gdaniec (2001: 190) erkennen: „Passive voice plays a role in creating the right focus in a sentence, among other things.“ Sie empfehlen gleichzeitig, das Passiv zum Zweck einer besseren maschinellen Übersetzbarkeit zu vermeiden, wenn es aus stilistischer Sicht nicht notwendig sei (ebd.). Nach Bernth und Gdaniec (ebd.: 191)

[25] Regel „Avoid post-modifying adjective phrases” bei Bernth & Gdaniec (2001), die auch von Bernth und Gdaniec für Deutsch empfohlen wurde.

[26] Die MÜ-Qualität wurde auf einer Skala von nn bis 3, in der nn für „Verbesserung nicht erforderlich“; 1 für „keine Verbesserung“; 2 für „leichte Verbesserung“; 3 für „starke Verbesserung“ stehen, gemessen.

könne der MÜ das Passiv aus zwei Gründen schwierig fallen: „It can be very difficult to disambiguate between stative and dynamic passives […]; (t)he argument assignments in passive constructions may differ between source and target language." Aufgrund der Notwendigkeit beider Formen untersucht die Studie, inwiefern der bisher realisierte MÜ-Fortschritt eine korrekte Übersetzung des Passivs ermöglicht.

4.5.2.3.7 Regel „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden"

Die Konstruktion „sein + zu + Infinitiv" ist im Deutschen eine Form des Passiversatzes für ein Passiv mit Modalverb (Beispiele: ‚Die Regeln müssen beachtet werden' → ‚Die Regeln sind zu beachten'; ‚Die zweite Aufgabe soll zunächst bearbeitet werden' → ‚Die zweite Aufgabe ist zunächst zu bearbeiten'), vgl. Teich (2003: 92). Ins Englische wird diese Form des Passivs als reguläres Passiv übersetzt (König & Gast 2012: 161) (‚The rules must be followed'; ‚The rules are to be followed'). Hierbei kann der kontrastive Konstruktionsunterschied zwischen dem Deutschen und Englischen (vgl. Teich 2003: 93) zu MÜ-Problemen führen. Reuther (2003) argumentiert, dass die systemspezifischen Eigenschaften die MÜ solcher stilistischen Phänomene beeinflussen. Darüber hinaus wird der Leser mit dem Passiversatz nicht direkt angesprochen, daher wird diese Regel (eine Formulierung mithilfe des Imperativs) von der tekom (2013: 28) für „eine schnelle und richtige Handlungsumsetzung" und von Congree (2018) für eine präzise und deutliche Anweisung empfohlen. Die Studie untersucht, inwiefern die MÜ-Systeme nach dem jetzigen Entwicklungsstand die Passiversatzkonstruktion fehlerfrei übersetzen können und wie diese Regel sich auf die MÜ-Qualität, vor allem die Stilqualität, auswirkt.

4.5.2.3.8 Regel „Überflüssige Präfixe vermeiden"

Bei dieser Regel wurde im Rahmen der Studie sowohl trennbare als auch untrennbare Verben[27] mit überflüssigen Präfixen analysiert. Untrennbare Verben sollten, solange sie ein Äquivalent im Englischen haben, unproblematisch übersetzt werden. Bei trennbaren Verben wird davon ausgegangen, dass sie – je nach Komplexitätsgrad des Satzes – den MÜ-Systemen Schwierigkeiten beim Parsen bereiten, wenn sie in der getrennten Form im Satz vorkommen. Dies ist z. B. der Fall, wenn das trennbare Verb ein Bestandteil eines Hauptsatzes (ohne Modalverb) ist; das

[27] Die analysierten Sätze enthalten 15 Sätze, in denen das Präfix getrennt am Satzende bzw. Nebensatzende erscheint sowie 9 Sätze mit untrennbaren Verben oder trennbaren Verben, in denen das Präfix ungetrennt vom Verb erscheint.

Präfix erscheint am Satzende getrennt vom Verb (Bsp. ‚Speichern Sie die angezeigten Werte lokal auf der Festplatte ab'). Im Englischen sind die ‚phrasal verbs' ebenfalls Verben, die aus einem Verb und einem Partikel bestehen. Allerdings ist die Satzstruktur im Englischen anders. Im Englischen kann das Partikel in manchen Fällen am Satzende platziert werden, in anderen Fällen hingegen ist dies nicht korrekt.[28]

Nicht nur die tekom (2013: 111) empfiehlt diese Regel zur Vereinfachung des Satzes und Reduzierung der Segmentvarianten bei der MÜ; diese Regel ist auch in mehreren KS-Regelsätzen enthalten. Siegel (2011) findet, dass die Verwendung trennbarer Verben in ihrer getrennten Form die Komplexität der Satzstruktur erhöhe. Bernth & Gdaniec (2001) führen diese Regel zur Reduzierung der Ambiguität und damit Verbesserung der maschinellen Übersetzbarkeit. Siegel (2013) zeigt, dass die Regelanwendung mit einer Verbesserung der MÜ-Qualität verbunden sei. Aufgrund der dargestellten sprachlichen Unterschiede und je nach Satzlänge und -komplexität könnten die Systeme Schwierigkeiten haben, das Präfix als Teil des Verbs zu erkennen. In der Studie wird die Anwendung dieser Regel unter die Lupe genommen, um zu untersuchen, inwiefern sie nach der heutigen Entwicklung der MÜ zur Verbesserung der maschinellen Übersetzbarkeit erforderlich ist.

4.5.2.3.9 Regel „Keine Wortteile weglassen"

Das Weglassen von Wortteilen ist eine Form der elliptischen Konstruktionen. Ellipsen werden von Halliday & Hasan (1976 zit. nach Bernth & Gdaniec 2001: 189) wie folgt definiert: „omission of something in the text, with the condition that what is omitted (or ellipted) is presupposed". Ziel des Weglassens von Wortteilen in beiden Sprachen ist die Reduzierung bzw. die Sprachökonomie (vgl. „Quantitätsmaxime" von Grice 1975: 26)[29] sowie die Erhöhung der Lesbarkeit (Bernth & Gdaniec 2001). Durch die Beschreibung der Ellipsen von Halliday und Hasan (ebd.: 189) als "specific structural slots to be filled by elsewhere" wird die Problematik ihrer MÜ ersichtlich. Die MÜ-Systeme verfügen nicht immer über Quellen zum Füllen der fehlenden Slots, daher führen die tekom (2013: 68) sowie mehrere weitere Regelsätze (Bernth & Gdaniec 2001; Reuther 2003; Siegel 2011; Congree 2018) diese Regel zur Unterstützung der maschinellen Übersetzbarkeit, Reduzierung der Ambiguität sowie ggf. Erhöhung der Textverständlichkeit auf.

[28] Während sowohl ‚She turned off the light' als auch ‚She turned the light off' richtig sind, ist nur ‚She looked after her brother' korrekt (‚She looked her brother after' wäre falsch).

[29] Die Quantitätsmaxime von Grice (1975: 26) bezieht sich auf die Quantität der Information: „Make your contribution as informative as is required (for the current purposes of the exchange) […] not more informative than is required".

Bei dem Sprachenpaar Deutsch – Englisch können die unterschiedlichen Rechtschreibregeln in Zusammenhang mit dieser Regel eine Rolle spielen: Im Deutschen wird ein Bindestrich bei der Abkürzung sowohl von Substantiven (z. B. ‚Vor- und Nachteile') als auch von Adjektiven (z. B. ‚prozess- und produktabhängig') verwendet. Im Englischen hingegen kann der Bindestrich bei der Abkürzung von Adjektiven (z. B. ‚process- and product-oriented'), aber nicht bei der Abkürzung von Substantiven verwendet werden. Aufgrund dieser orthografischen Unterschiede und dadurch, dass bei der Anwendung dieser Regel die Verwendung des Bindestriches entfällt, wird erwartet, dass die Regel die MÜ-Systeme bei der Tokenisierung unterstützt. Des Weiteren kann der Geläufigkeitsgrad der abgekürzten Begriffe eine besondere Rolle bei der Korrektheit ihrer MÜ spielen.[30] Es wird davon ausgegangen, dass gebräuchliche abgekürzte Begriffe in ihrer abgekürzten Form in den MÜ-Systemen lexikalisch hinterlegt sind bzw. in den Trainingsdaten vorkommen. Dementsprechend wäre in dem Fall die Wahrscheinlichkeit einer korrekten Übersetzung hoch und die Regelanwendung – zum Zwecke der maschinellen Übersetzbarkeit – nicht erforderlich. Die Regelanwendung kann sogar bei Begriffen, die in ihrer abgekürzten Form gängig sind (z. B. ‚Gebrauchs- und Bedienungs*anleitung*' → ‚Gebrauchs*anleitung* und Bedienungs*anleitung*'), stilistisch unakzeptabel sein, was berücksichtigt werden muss, wenn der Stil bei der entsprechenden Textsorte Vorrang haben sollte. Bei ungebräuchlichen abgekürzten Begriffen hingegen könnte die Verwendung der vollständigen Wörter (d. h. eine Anwendung der Regel) eine korrekte MÜ fördern, wie es in den obengenannten Regelsätzen beabsichtigt ist. Die Angabe von vollständigen Wörtern kann einerseits insbesondere bei kritischen Kontexten (z. B. Sicherheitshinweise) inhaltlich für höhere Klarheit sorgen. Auf der anderen Seite kann ebenfalls hier die Wiederholung von Wörtern (wie ‚Konfiguration' in ‚Eingangskonfiguration und Ausgangskonfiguration' (input configuration and output configuration)) im Deutschen sowie im Englischen stilistisch kritisch betrachtet werden. Daher wird bei dieser Regel typischerweise dem Autor ein gewisser Freiraum für die Entscheidung über die Regelanwendung je nach den gegebenen Umständen eingeräumt (Bernth & Gdaniec 2001). Vor diesem Hintergrund ist es von Interesse, in der Studie zu untersuchen, ob in der Zwischenzeit die MÜ dieser Art der Ellipsen sich nach jetzigem Entwicklungsstand der Systeme verbessert hat, wodurch von einem prägnanten Stil profitiert werden kann.

[30] Da die Studie generische Black-Box-Systeme untersuchte, bei denen keine Terminologieintegration erfolgen konnte, wurden firmen- und produktspezifische Termini durch geläufige Begriffe ersetzt. Für die Auswahlkriterien der untersuchten Systeme siehe §4.5.1. Für den genauen Umgang mit den spezifischen Termini im Rahmen der Studie siehe Schritt [4] unter §4.5.3.1.

4.5.3 Datensatz: Beschreibung und Aufbereitung

In diesem Abschnitt wird der Studiendatensatz präsentiert. Nach einer zusammenfassenden Beschreibung des Datensatzes werden der Aufbau des Datensatzes sowie die Aufbereitungsschritte für die Analysemethoden genauer erläutert.

Kurze Beschreibung des Datensatzes: Bei den analysierten Daten handelt es sich um eine korpusbasierte Testsuite. 216 Ausgangssätze wurden aus einem Korpus von zehn deutschen Benutzerhandbüchern für Geräte, Software und Maschinen extrahiert. Diese Ausgangssätze verstoßen gegen mindestens eine der neun untersuchten KS-Regeln. Die KS-Regeln wurden einzeln in den jeweiligen Sätzen angewendet. Somit entstanden zwei Versionen von jedem Satz: eine Version vor dem Einsatz der KS-Regeln (nachstehend „vor-KS“ genannt) und eine Version nach dem Einsatz der KS-Regeln (nachstehend „nach-KS“ genannt). Beide Versionen wurden von den fünf MÜ-Systemen ins Englische übersetzt. Entsprechend wurde ein Datensatz aus 2.160 MÜ-Sätzen (216 Ausgangssätze * 2 Versionen * 5 Systeme) gebildet. Dieser Datensatz wurde komplett im Rahmen der Fehlerannotation analysiert. 1.100 MÜ-Sätze von den 2.160 wurden im Rahmen der Humanevaluation bewertet.

4.5.3.1 Details der Erstellung und Aufbereitung des Datensatzes

Zur Untersuchung des Einflusses der *einzelnen* KS-Regeln musste die Analyse auf Satzebene erfolgen. Wie unter §3.4.2 „Evaluationsdesign“ erklärt, wurde entschieden, die Studie auf Basis einer korpusbasierten Testsuite (vgl. Balkan & Fouvry 1995) durchzuführen, um von der Textauthentizität und sprachlichen Varianz eines Textkorpus (vgl. Engelberg 2009) zu profitieren und gleichzeitig eine gewisse Steuerung über die Datenkonstruktion beizubehalten. Eine gewisse Steuerung des Datensatzes war erforderlich, um die Komplexität eines natürlichen Texts – nach klar definierten Kriterien – in einem Rahmen zu reduzieren, der der Analyse der KS in Zusammenhang mit der MÜ dienlich ist. Die Vorgehensweise zur Erstellung und Aufbereitung des Datensatzes erstreckt sich über die folgenden zehn Schritte:

[1] Erstellung eines einsprachigen Korpus

[2] Prüfung der Konformität mit den analysierten KS-Regeln

[3] Auswahl der Testsätze (Ausgangssätze)

[4] Aufbereitung der Testsätze (Ausgangssätze)

[5] Einsatz der KS-Regeln

[6] Qualitätsprüfung der Ausgangssätze (DE)

[7] Übersetzung und Fehlerannotation

[8] Auswahl der MÜ-Sätze (Zielsätze) für die Humanevaluation

[9] Aufbereitung der MÜ-Sätze (Zielsätze) für die Humanevaluation

[10] Qualitätsprüfung der Zielsätze (EN)

Im Folgenden werden die Vorgehensweise zusammen mit einer genauen Erläuterung der Auswahlkriterien sowie die Qualitätsprüfungsverfahren des Datensatzes detaillierter dargestellt:

[1] Erstellung eines einsprachigen Korpus: Es wurde ein einsprachiges Korpus aus zehn deutschen Betriebsanweisungen, Pflegeanleitungen, Regelungen sowie Benutzerhandbüchern von Haushaltsgeräten, Software und Maschinen verschiedener Hersteller erstellt. Die verschiedenen Dokumente wurden online als PDF-Dateien heruntergeladen und in Textdateien konvertiert. Die Dateien wurden durch das Entfernen sämtlicher Grafiken, Tabellen, Auflistungen und Überschriften bereinigt. Die Absätze wurden in einzelne Sätze zerlegt und aufgelistet.

[2] Prüfung der Konformität mit den analysierten KS-Regeln: Die Textdateien wurden in CLAT auf Verstöße gegen die neun analysierten KS-Regeln[31] untersucht. Sätze, die Verstöße aufweisen, wurden in neun Excel-Tabellen extrahiert. Jede Excel-Tabelle beinhaltet die Sätze mit den Verstößen gegen eine bestimmte Regel unter Angabe der Quelle des Satzes sowie der mithilfe von Excel ermittelten Satzlänge.

[3] Auswahl der Testsätze (Ausgangssätze): Zum Filtern der Testsätze wurden sie in einem *einleitenden Schritt mit den fünf* verwendeten MÜ-Systemen übersetzt. Unter Berücksichtigung einer möglichst ausgewogenen Auswahl aus allen Quellen (d. h. Benutzerhandbüchern) wurden die ersten enthaltenen Sätze ausgewählt, die die folgenden Kriterien erfüllen:

[31] Verstöße gegen einige Regeln wie „Funktionsverbgefüge vermeiden“ konnten von CLAT nicht immer erkannt werden (meistens werden nur bekannte FVG wie ‚erfolgen‘ oder ‚vornehmen‘ gut erkannt), daher wurden die Sätze in manchen Fällen manuell extrahiert.

- Die maschinelle Übersetzung von mehr als zwei MÜ-Systemen ist akzeptabel: Die Sätze wurden in den Datensatz aufgenommen, wenn die MÜ grundsätzlich akzeptabel war. In anderen Worten wurden die Sätze, deren MÜ mit den meisten Systemen unverständlich waren, ausgeschlossen. Da der Fokus der Studie darin besteht, den Einfluss einzelner KS-Regeln auf die MÜ genauer zu analysieren (und nicht den MÜ-Output im Allgemeinen zu beurteilen), ist die Untersuchung einer komplett unverständlichen MÜ nicht zielführend.

- Die MÜ von mindestens zwei MÜ-Systemen beinhaltet mindestens einen Fehler: Es wurden die Sätze ausgeschlossen, die sowohl vor-KS als auch nach-KS von allen Systemen fehlerfrei übersetzt wurden. Diese sind in den meisten Fällen technische Standardsätze oder AGB, die aufgrund ihres häufigen Auftretens insbesondere von SMÜ-Systemen fehlerfrei übersetzt werden.

- Die Sätze ergeben alleinstehend, sprich ohne weitere Kontextinformation, Sinn, damit die Teilnehmer der Humanevaluation die Satzsemantik der Ausgangssätze ohne Kontext erkennen und entsprechend die MÜ bewerten können.

Zum Schluss wurden 24 Sätze bei jeder KS-Regel ausgewählt (insgesamt 216 Sätze). Somit liegt die Anzahl der analysierten Sätze pro Regel über der einschlägiger Studien. Roturier (2006) analysierte zwei bis 14 Sätze pro Regel (insgesamt 54 Regeln; 304 Sätze). O'Brien (2006) untersuchte insgesamt 130 Segmente bei 29 NTIs (Negative Translatability Indicators), d. h. durchschnittlich 4,5 Segmente pro NTI, wobei die Anzahl der Segmente pro NTI variierte. In Dohertys Korpus wurden insgesamt 33 Verstöße gegen zehn KS-Regeln identifiziert, wobei die Anzahl der Verstöße bei den einzelnen Regeln unterschiedlich ausfiel und zwischen einem und zwölf Verstößen lag (Doherty 2012: 103).

Im Rahmen dieser Studie schafft die Analyse einer gleichen Anzahl an Sätzen pro Regel sowohl bei der quantitativen als auch bei der qualitativen Analyse eine einheitliche Untersuchungsbasis. Eine weitere Erhöhung der Anzahl der Sätze könnte die Integrität der Ergebnisse gefährden. MÜ-Evaluationsstudien thematisieren die Abwägung zwischen der Größe des Datensatzes und der Integrität der Ergebnisse (Fiederer & O'Brien 2009). Coughlin (2003: 2) analysierte 124 MÜ-Sätze und machte die Beobachtung: „Rating translation quality is both tedious and repetitive". Es kam vor (ebd.), dass die Bewerter identischen Sätzen unterschiedliche Scores gaben oder Sätze, die mit der Referenzübersetzung

identisch waren, nicht mit dem besten Score bewerteten. In einer weiteren MÜ-Evaluationsstudie wurde der Datensatz auf 180 Sätze begrenzt, um das „risk of boredom and its negative consequences such as inconsistent evaluation or diminishing powers of judgement“ zu minimieren (Fiederer & O'Brien 2009: 59). In Tabelle 4.2 sind die analysierten Dokumente sowie die Aufteilung der Sätze genauer angegeben:

Wie Tabelle 4.2 zeigt, steht der Anteil der analysierten Sätze im Verhältnis zu der Größe der jeweiligen Quelle.

[4] Aufbereitung der Testsätze (Ausgangssätze): Nach King & Falkedal (1990) „[t]he major problem with setting up test suites of this kind is that of interaction between different linguistic phenomena. The standard solution is to reduce the linguistic complexity of all items other than the item of direct interest to an absolute minimum". Dementsprechend wurden die ausgewählten Sätze wie folgt aufbereitet:

Orthografie

- *Eigennamen bzw. Bezeichnungen* wurden großgeschrieben (z. B. ‚Manueller' in ‚manueller Betrieb' im Satz ‚Die Funktion manueller Betrieb soll ausgewählt werden').
- *Rechtschreibfehler* wurden korrigiert, z. B. wurden überflüssige Kommas gelöscht.

Lexik

- *Anonymisierung der Sätze*: Firmen- und Produktnamen wurden z. B. durch ‚die Firma', ‚das Programm', ‚das Gerät' usw. ersetzt.
- *Überflüssige Details*, Beispiele, Aufzählungen wurden eliminiert, um die Satzlänge zu kürzen.
- *Füllwörter und Adverbien* wurden eliminiert, wenn bei mehr als zwei MÜ-Systemen ersichtlich war, dass sie für die MÜ problematisch sind. Das galt für Adverbien, solange sie für die Satzsemantik nicht zwingend notwendig waren.

[32] Die Forscherin richtet ihren Dank an alle Firmen, die zum Forschungszweck ihre Dokumente zur Verfügung gestellt haben. Ohne diese Dokumentation wäre die Studie nicht zustande gekommen. Auf Wunsch der meisten Firmen bleiben die Herstellernamen anonym.

Tabelle 4.2: Anteil der analysierten Sätze im Verhältnis zu der Größe der jeweiligen Quelle

Dokument[32]	Anzahl der Wörter des Dokuments	Anzahl der Sätze des Dokuments	% im Korpus	Anzahl der analysierten Sätze	% im Datensatz
Verlegeanweisung und Reinigungsverfahren eines Teppichbodens	2.263	143	6,5 %	18	8,3 %
Gebrauchs- und Pflegeanleitung für keramikversiegeltes Kochgeschirr sowie AGB	1.198	74	3,4 %	10	4,6 %
Bedienungsanleitung eines Feinstzerkleinerers (ein Schneidesystem zum Zerkleinern von Lebensmittel-, Chemie- bzw. Medizinprodukten)	5.851	494	22,4 %	35	16,2 %
Handbuch einer Software zur Parametrierung eines Volumenstrom-Kompaktreglers (der Volumenstrom-Kompaktregler ist für Labor- und Pharmaanwendungen bestimmt)	9.776	678	30,7 %	92	42,6 %
Gepäckregelung einer Fluggesellschaft	913	53	2,4 %	4	1,9 %
Bedienungsanleitung eines Milchaufschäumers	1.310	94	4,3 %	9	4,2 %
Pflege- und Bedienungsanleitung eines Küchenmöbels	3.459	223	10,1 %	15	6,9 %
Technische Beschreibung und Bedienungsanleitung eines Heimkino-Lautsprecher-Sets	2.065	156	7,1 %	14	6,5 %
Bedienungsanleitung einer elektrischen Zitruspresse (Elektro-Hausgerät)	1.031	82	3,7 %	12	5,6 %
Betriebsanweisung einer Serie von Haus- und Gartenpumpen	2.154	210	9,5 %	7	3,3 %
	30020	**2207**	**100 %**	**216**	**100 %**

- *Firmen- sowie produktspezifische Termini und ungebräuchliche Fachtermini* wurden durch geläufige Begriffe ersetzt, z. B. wurde ‚Gerät' anstatt ‚Feinstzerkleinerer', ‚Steckdose' anstatt ‚Schutzkontaktsteckdose' verwendet. Ein Terminus wurde ersetzt, wenn er in den zwei häufig verwendeten Online-Wörterbüchern „dict.cc" oder „leo.org" nicht vorhanden war. Diese Anpassung wurde aus zwei Gründen vorgenommen; (1) damit sich die Bewerter während der Humanevaluation nicht mit der Suche nach genauen Übersetzungen für solche Termini beschäftigen; (2) aufgrund der Spezifität solcher Termini ist nicht zu erwarten, dass sie von den untersuchten generischen MÜ-Systemen korrekt übersetzt werden (siehe Auswahlkriterien der Systeme unter §4.5.1). In der Regel führen die Unternehmen Teminologiedatenbanken und integrieren sie in ihre implementierten Übersetzungsprogramme oder MÜ-Systeme. Die Terminologieintegration wäre zur Zeit der Durchführung der empirischen Studie (Ende 2016 – Anfang 2017) zwar in der RMBÜ, SMÜ und HMÜ, jedoch nicht in der NMÜ, möglich gewesen (vgl. Eisold 2017). Damit die Studie auf einer einheitlichen Basis durchgeführt wird, wurden alle Systeme in ihrem Ist-Zustand, d. h. ohne Terminologieintegration oder Training mit domänenspezifischen Daten, verwendet.[33]

  Das Ersetzen solcher seltenen Termini wurde bei allen analysierten KS-Regeln mit Ausnahme der Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" vorgenommen, da Oberflächentexte und Bedienelemente naturgemäß firmen-, produktspezifische bzw. ungebräuchliche Termini beinhalten. Durch diese Ausnahme konnte der Effekt der Verwendung der geraden Anführungszeichen auf die Übersetzung von seltenen Termini bei den unterschiedlichen Systemen getestet und untersucht werden.

- Sätze mit mehr als einem spezifischen bzw. ungebräuchlichen Wort wurden von der Analyse ausgeschlossen, um zu vermeiden, dass mehrere Änderungen im Satz vorgenommen werden.

Grammatik

- *Satzstruktur* wurde nicht verändert.

[33] Mehr zur Entwicklung der NMÜ im Bereich der Domänenadaptation bzw. Terminologieintegration unter §3.3.4.

Satzlänge bei den einzelnen Regeln Die durchschnittliche Satzlänge betrug regelübergreifend 12 Wörter. Die Satzlänge variierte zwischen 5 und 17 Wörtern, wobei 97,2 % der 216 Sätze eine Länge zwischen 8 und 16 Wörtern[34] aufwiesen. Eine Spanne von 8 Wörtern ist relativ groß. Dies war jedoch nicht zu vermeiden, da bei manchen Regeln, wie z. B. der Regel „Eindeutige pronominale Bezüge verwenden", zwangsläufig ein Satz mit einem Haupt- und Nebensatz erforderlich ist. Eine Untergrenze der Satzlänge wurde im Vorfeld nicht gesetzt. Maßgeblich war, dass der Satz vollständig ist (d. h. Phrasen wurden ausgeschlossen). Längere Sätze (über 17 Wörter) wurden aus zwei Gründen ausgeschlossen: (1) Obwohl die meisten KS-Regeln eine Satzlänge bis 25 Wörter erlauben (vgl. Wells Akis u. a. 2003; O'Brien 2003; Aikawa u. a. 2007; Fiederer & O'Brien 2009; Mügge 2013), so steigt mit der Länge der analysierten Sätze, auch die Wahrscheinlichkeit, dass die Bewerter während der Humanevaluation von Teilen außerhalb der KS-Stelle abgelenkt werden. (2) Eine Aufbereitung von langen oder komplexen Sätzen nach der Vorgehensweise unter Schritt [4] wäre mit mehreren Anpassungen und ggf. mehreren Bearbeitungsmöglichkeiten verbunden. Hierfür können keine klar definierten Bearbeitungsregeln festgelegt werden. Die Länge der Sätze bei den einzelnen Regeln sah wie folgt aus (Tabelle 4.3):

Tabelle 4.3: Länge der Sätze bei den einzelnen KS-Regeln

	anz	fvg	kos	nsp	pak	pas	per	prä	wte
Mittelwert	11	11	13	14	11	11	10	12	12
Max	15	15	15	16	15	15	15	16	17
Min	8	8	8	8	7	8	5	8	7
Median	11	11	13	15	11	11	11	11	12
Standardabw.	2,12	2,05	1,89	2,33	1,86	1,90	2,71	2,33	2,54

anz: Für zitierte Oberflächentexte gerade Anführungszeichen verwenden
fvg: Funktionsverbgefüge vermeiden
kos: Konditionalsätze mit ‚Wenn' einleiten
nsp: Eindeutige pronominale Bezüge verwenden
pak: Partizipial-konstruktionen vermeiden
pas: Passiv vermeiden
per: Konstruktionen mit „sein + zu + Infinitiv" vermeiden
prä: Überflüssige Präfixe vermeiden
wte: Keine Wortteile weglassen

[34] Nur bei 6 Sätzen lag die Länge bei 5 Wörtern (1 Satz), 6 Wörtern (2 Sätze), 7 Wörtern (2 Sätze) und 17 Wörtern (1 Satz).

[5] Einsatz der KS-Regeln Nachdem die Ausgangssätze, die gegen die KS-Regeln verstoßen, im vorherigen Aufbereitungsschritt festgelegt wurden (Version vor-KS), erfolgte eine Umsetzung der neun KS-Regeln. Sollte ein Satz Verstöße gegen mehr als eine KS-Regel enthalten, wurde nur der für die untersuchte Regel relevante Verstoß eliminiert, denn die Anwendung mehrerer Regeln würde eine Abgrenzung der Auswirkungen jeder Regel erschweren. Zudem ist eine Korrektur verschiedener Verstöße in einem Satz nicht immer möglich. Vorherige Studien thematisierten diese Herausforderungen, die mit der Untersuchung des Einflusses *einzelner* KS-Regeln einhergehen (vgl. Nyberg u. a. 2003; Doherty 2012: 30; Ramírez Polo 2012: 274).[35] Ramirez Polo (ebd.: 274) betonte, dass „it is not always straightforward to determine which rule had an effect on the quality of the segment, since many segments were affected by more that one rule". Roturier (2006: 74) wendete ebenfalls nur die untersuchte Regel im Satz an, um zu vermeiden, dass „some of the test suite's segments could be contaminated and the internal validity of the study would be undermined" (ebd.). Das Ergebnis dieses Schrittes ist der Erhalt einer zweiten Version von jedem Ausgangssatz, die KS-regelkonform ist (Version nach-KS).

[6] Qualitätsprüfung der Ausgangssätze (DE) Vor der Durchführung der maschinellen Übersetzung war es zunächst erforderlich, die Qualität der Ausgangssätze zu prüfen, da die Qualität des Ausgangstexts einen unmittelbaren Einfluss auf die des MÜ-Outputs hat. Mehrere Studien betonten die Bedeutung der Ausgangstextqualität bei der MÜ-Evaluation (vgl. Brunette 2000; Castilho & O'Brien 2016). In der Regel wird der Fokus bei der Evaluation auf den Zieltext als das Endprodukt gelegt, wobei der Ausgangstext außer Acht gelassen wird (vgl. Molnár 2012).

In der vorliegenden Studie wurde die Qualität der Ausgangssätze wie folgt geprüft: Zuerst wurden die vor-KS- und nach-KS-Ausgangssätze von einer erfahrenen (7 Jahre Berufserfahrung) Übersetzerin mit Deutsch als Muttersprache orthografisch und grammatisch geprüft. In einem zweiten Schritt wurden die Ausgangsätze auf stilistische Akzeptanz geprüft. Die Prüfung wurde von zwei deutschen Linguisten durchgeführt: Der erste ist ein Akademiker im Bereich der Translationswissenschaft, Lexikograf und Autor von sechs Wörterbüchern. Der zweite ist Universitätsprofessor für Übersetzungswissenschaft mit dem Schwerpunkt MÜ und technische Übersetzung. Diese Prüfung zielte darauf ab, stilistisch kritische Sätze zu identifizieren. Stilistisch kritisch ist ein Satz dann, wenn er von einem deutschen Muttersprachler auf die vorhandene Weise nicht formuliert

[35] Eine Darstellung der Forschungsherausforderungen ist unter §3.5.3 enthalten.

werden würde (d. h. nicht authentisch klingt). Bei der Prüfung sollte zunächst der erste Beurteiler bei jedem Satz ein Häkchen setzen, um zu signalisieren, ob er den Satz stilistisch akzeptabel findet oder nicht. Sätze, die als stilistisch kritisch beurteilt wurden, durchliefen eine erneute Prüfung durch den zweiten Beurteiler.

Von den insgesamt 216 Ausgangssätzen wurden 27 Sätze (12,5 %) nach-KS von dem ersten Beurteiler als stilistisch kritisch eingestuft. Beispielsweise war bei der Regel „Partizipialkonstruktionen vermeiden" die Formulierung nach-KS des folgenden Satzes umstritten:

Vor-KS: *Das Gerät verbindet sich mit der neu gewählten Netzwerkadresse.*

Nach-KS: *Das Gerät verbindet sich mit der Netzwerkadresse, die neu gewählt wird.*

12 der 27 Sätze wurden ebenfalls vom zweiten Beurteiler als stilistisch kritisch klassifiziert. Diese Sätze wurden durch neue Sätze ersetzt und nach demselben Ablauf geprüft. Es blieben 15 Sätze (6,9 % der 216 Ausgangssätze) der stilistisch kritischen Sätze umstritten. Die 15 Sätze kamen bei drei der neun KS-Regeln vor: vier Sätze in „Partizipialkonstruktionen vermeiden", fünf Sätze in „Eindeutige pronominale Bezüge verwenden" und sechs Sätze in „Keine Wortteile weglassen".[36]

[7] Übersetzung und Fehlerannotation Die Ausgangssätze vor und nach dem Einsatz der KS-Regeln wurden mit den fünf MÜ-Systemen ins Englische übersetzt. Bei der Übersetzung der zwei Versionen der 216 Sätze mit fünf MÜ-Systemen ergibt sich ein Datensatz von 2.160 MÜ-Sätzen. Die Übersetzungsfehler in den MÜ-Sätzen wurden nach der festgelegten Fehlertaxonomie von einem in DE-EN vereidigten Übersetzer mit sechs Jahren Berufserfahrung annotiert (siehe §4.5.4.1 und §4.5.4.2). Daraufhin wurden die annotierten Fehler von zwei professionellen Linguisten geprüft. Aufgrund der großen Anzahl der MÜ-Sätze (2.160 Sätze) prüfte jeder Linguist nur die Hälfte der MÜ-Sätze. Die Prüfung fand in Form einer Bewertung jedes Satzes statt, in der der Prüfer ankreuzen sollte, ob er der Annotation zustimmt oder nicht. Im Fall, dass er der vorherigen Annotation nicht zustimmte, musste er die Übersetzung annotieren. Der zweite Linguist prüfte im Anschluss beide Annotationen und entschied sich für eine davon.

Anschließend wurde die KS-Stelle nach den Regeln in §4.5.2.2 identifiziert. Die Identifizierung der KS-Stelle in der MÜ erfolgte manuell, wie die Definition der

[36] Bei der statistischen Analyse zeigten diese stilistisch kritischen Sätze keinen signifikanten Unterschied im Vergleich zu dem Rest der Sätze im Hinblick auf die Fehleranzahl- und die Qualitätsveränderung.

KS-Stelle besagt, als Äquivalent der durch den Einsatz der KS-Regel umformulierten Stelle im Ausgangssatz. Eine automatische Identifizierung der MÜ einer bestimmten Stelle im Ausgangssatz auf Basis von Techniken des Wort-Alignments war laut vorherigen Studien in manchen Fällen ungenau bzw. fehlerhaft (vgl. Koehn u. a. 2003; Och & Ney 2004; Callison-Burch u. a. 2007), daher hat die Forscherin diesen Schritt manuell durchgeführt.

[8] Auswahl der MÜ-Sätze (Zielsätze) für die Humanevaluation Nicht alle 2.160 annotierten MÜ-Sätze wurden in der Humanevaluation bewertet. Dies ist auf zwei Gründe zurückzuführen:

Erstens – da jeder Ausgangssatz in zwei Versionen (vor-KS und nach-KS) von fünf Systemen übersetzt wurde, existierten für jeden Ausgangssatz 10 MÜ. Die 10 Übersetzungen sind in vielen Fällen relativ ähnlich. Dementsprechend war es schwer vorstellbar, dass die Teilnehmer sie trotz Eintönigkeit konzentriert bzw. ohne negativen Einfluss auf ihre Leistung bewerten könnten. Um diesem Risiko entgegenzuwirken, war es erforderlich, die Anzahl der MÜ auf 5–6 pro Ausgangssatz einzugrenzen.

Zweitens – Um das Studienziel (Untersuchung des Einflusses der *einzelnen* KS-Regeln) zu realisieren, war es erforderlich, dass die Bewerter in der Humanevaluation möglichst nur die KS-Stelle[37] im Fokus behielten. Dementsprechend mussten alle Fehler außerhalb der KS-Stelle korrigiert werden. Dieser grundlegende Schritt war unerlässlich, andernfalls hätten die Bewerter die außerhalb der KS-Stelle aufgetretenen Fehler bei der Bewertung mitberücksichtigt. Nicht alle MÜ-Sätze lassen sich außerhalb der KS-Stelle korrigieren, *ohne dass diese Korrektur die KS-Stelle beeinflusst.*

In der Regel sind Übersetzungen, die mehr als 2-3 falsche Wörter[38] beinhalten, schwer zu korrigieren,[39] ohne dass die KS-Stelle von der Korrektur beeinflusst werden würde. Dieser Grenzbereich stellte das Ausschlusskriterium bei der Auswahl der Zielsätze für die Humanevaluation dar. Das Ausschlusskriterium wurde folgendermaßen umgesetzt:

- *Im Falle der Übersetzungen, die innerhalb der KS-Stelle Fehler beinhalten,* wurden die Sätze analysiert, in denen max. 2 Wörter *außerhalb* der KS-Stelle Fehler beinhalten. Das Ausschlusskriterium wurde unabhängig von

[37] Mehr zur „KS-Stelle“ unter §4.5.2.1.

[38] Ein falsches Wort kann einen oder mehrere Fehler aufweisen, z. B. Großschreibfehler und falsche Wortstellung.

[39] Eine genaue Beschreibung der Aufbereitung der MÜ-Sätze ist unter Schritt [9] aufgeführt.

der Fehleranzahl *innerhalb* der KS-Stelle umgesetzt, vorausgesetzt die Fehler außerhalb der KS-Stelle konnten ohne Einfluss auf die KS-Stelle korrigiert werden.

- *Im Falle der Übersetzungen, die nur außerhalb der KS-Stelle Fehler beinhalten,* wurden die Sätze analysiert, in denen max. 3 Wörter Fehler beinhalten, solange die Korrektur der Fehler außerhalb der KS-Stelle keinen Einfluss auf die KS-Stelle hat.

In einem zweiten Schritt wurde überprüft, ob die bei jeder Regel häufig vorgekommenen Fehlertypen durch das Ausschlusskriterium nicht reduziert wurden. Sollte ein häufig vorgekommener Fehlertyp nicht in einem ähnlichen Verhältnis repräsentiert sein, wurden die ausgeschlossenen Übersetzungen nochmal überprüft und die Anzahl der zulässigen falschen Wörter erhöht, damit eine Übersetzung mit diesem Fehlertyp miteinbezogen werden konnte. Die einzige Voraussetzung hierfür ist, dass diese MÜ außerhalb der KS-Stelle gut korrigierbar ist (d. h. ohne Einfluss auf die KS-Stelle zu haben), z. B. befinden sich die Fehler im Hauptsatz und die KS-Stelle im Nebensatz.

[9] Aufbereitung der MÜ-Sätze (Zielsätze) für die Humanevaluation Nachdem die Zielsätze nach den genannten Kriterien herausgefiltert wurden, wurden sie für die Humanevaluation aufbereitet. Die Aufbereitung bestand aus zwei Schritten: einer Korrektur der Fehler außerhalb der KS-Stelle sowie einer Vereinheitlichung der MÜ-Sätze außerhalb der KS-Stelle.

Fehler außerhalb der KS-Stelle wurden mit der minimalen Anzahl von Edits korrigiert. Tabelle 4.4 zeigt, wie die Fehler außerhalb der KS-Stelle korrigiert wurden:

In Tabelle 4.4 geht es um die KS-Regel „Funktionsverbgefüge vermeiden". Nur drei von den fünf untersuchten Systemen werden dargestellt, da die MÜ der zwei weiteren Systeme (Bing und Google Translate) identisch mit der MÜ von SDL war. In diesem Beispiel beinhaltete nur die MÜ von Lucy einen Fehler außerhalb der KS-Stelle (‚not purpose-appropriate'). Bei der Aufbereitung der MÜ-Sätze für die Humanevaluation wurde dieser Fehler korrigiert.

Der zweite aufbereitende Schritt war die Vereinheitlichung der MÜ aller MÜ-Systeme außerhalb der KS-Stelle. Dieser Schritt wurde aus zwei Gründen vorgenommen: Durch die Vereinheitlichung (1) lässt sich der Einfluss verschiedener Fehlertypen auf die MÜ desselben Ausgangssatzes vergleichen; und (2) können die MÜ der KS-Stelle in den einzelnen Systemen einander gegenübergestellt werden. Das Ergebnis dieses Schritts war im Falle des vorherigen Beispielsatzes wie folgt (Tabelle 4.5).

Tabelle 4.4: Beispiel 2

Vor KS	Der Hersteller **übernimmt** keine **Haftung** für Schäden, die durch nicht bestimmungsgemäßen Gebrauch entstanden sind.
SMÜ SDL	The manufacturer accepts no liability for damage caused by improper use.
HMÜ Systran	The manufacturer does not take over liability for damage, which resulted from not intended use.
RBMÜ Lucy	The manufacturer does not take over liability for damages which arose through use *not purpose-appropriate*.
Nach KS	Der Hersteller **haftet** nicht für Schäden, die durch nicht bestimmungsgemäßen Gebrauch entstanden sind.
SMÜ SDL	The manufacturer is not liable for damage caused by improper use.
HMÜ Systran	The manufacturer is not responsible for damage, which resulted from not intended use.
RBMÜ Lucy	The manufacturer is not liable for damages which arose through use *not purpose-appropriate*.

Die KS-Stelle ist farblich dargestellt: Blau wird für die korrekten Teile der Übersetzung verwendet; rot für die falschen Teile. Fehler *außerhalb* der KS-Stelle, die korrigiert wurden, sind *kursiv und unterstrichen* dargestellt.

Übersetzungen, die außerhalb der KS-Stelle in allen Systemen nicht vereinheitlicht werden können, wurden von der Humanevaluation ausgeschlossen, damit die Bewertungen aller MÜ jedes Ausgangssatzes vergleichbar bleiben.

Zwei Ausnahmen waren bei der Aufbereitung der MÜ-Sätze zu berücksichtigen: (1) Der Wortstellungsfehler wurde von der Korrektur außerhalb der KS-Stelle ausgenommen. Eine Korrektur der Wortstellung außerhalb der KS-Stelle würde dazu führen, dass die Wortstellung innerhalb der KS-Stelle geändert bzw. korrigiert wird. In diesem Fall wurden beide Wortstellungsfehler (innerhalb und außerhalb der KS-Stelle) beibehalten. Tabelle 4.6 zeigt diesen Fall:

In Tabelle 4.6 geht es um die KS-Regel „Passiv vermeiden“. In diesem Beispiel wurde der Wortstellungsfehler (in ‚a safety plug‘) nicht behoben, damit die Wortstellung der KS-Stelle nicht korrigiert wird. Der andere, semantische Fehler (‚equipment‘) außerhalb der KS-Stelle wurde korrigiert.

Tabelle 4.5: Beispiel 3

Vor KS	Der Hersteller **übernimmt** keine **Haftung** für Schäden, die durch nicht bestimmungsgemäßen Gebrauch entstanden sind.
SMÜ SDL	The manufacturer accepts no liability for any damage caused by improper use.
HMÜ Systran	The manufacturer does not take over liability for any damage caused by improper use.
RBMÜ Lucy	The manufacturer does not take over liability for any damage caused by improper use.
Nach KS	Der Hersteller **haftet** nicht für Schäden, die durch nicht bestimmungsgemäßen Gebrauch entstanden sind.
SMÜ SDL	The manufacturer is not liable for any damage caused by improper use.
HMÜ Systran	The manufacturer is not responsible for any damage caused by improper use.
RBMÜ Lucy	The manufacturer is not liable for any damage caused by improper use.

Hervorgehobene Stellen zeigen die vereinheitlichten Stellen in der MÜ auf.

Tabelle 4.6: Beispiel 4

Vor KS	Bei der Arbeit mit elektrischen Geräten **sollte** stets ein Sicherheitsstecker **verwendet werden.**
SMÜ SDL Fehlerannotation	When working with electrical *equipment* should always be ***a safety plug*** is used.
SMÜ SDL Humanevaluation	When working with electrical *devices*, should always be ***a safety plug*** is used.

Die KS-Stelle ist farblich dargestellt: Blau wird für die korrekten Teile der Übersetzung verwendet; rot für die falschen Teile. Fehler *außerhalb* der KS-Stelle, die korrigiert wurden, sind *kursiv und unterstrichen* dargestellt; bzw. ***fett, kursiv und unterstrichen*** bei Wortstellungsfehlern *außerhalb* der KS-Stelle, die nicht korrigiert wurden.

(2) Der zweite Sonderfall trat bei der Regel „eindeutige pronominale Bezüge verwenden“ auf. Bei dieser Regel spielt der Bezug bei der Bewertung eine Hauptrolle. Hierbei wäre eine semantische Vereinheitlichung aller MÜ außerhalb der KS-Stelle irreführend. Dieser Fall wird durch Tabelle 4.7 veranschaulicht. Die vor-KS Version lautete ‚Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, *ihn* rückstandslos zu entfernen‘:

Nach der Regel „eindeutige pronominale Bezüge verwenden“ wurde das Pronomen ‚ihn‘ durch seinen Bezug ‚den Fleck‘ in der nach-KS Version ersetzt. Das Wort ‚Fleck‘ wurde von den beiden Systemen unterschiedlich übersetzt (‚mark‘ bei Systran; ‚spot‘ bei Lucy). Entsprechend übersetzten die Systeme den Bezug *innerhalb* der KS-Stelle (blauer Text im Beispiel) unterschiedlich. Bei der Aufbereitung wurde die Übersetzung des Worts ‚Fleck‘ *außerhalb* der KS-Stelle (hervorgehoben im Beispiel) bei den unterschiedlichen Systemen *nicht* vereinheitlicht. Alle anderen unterschiedlichen Stellen in den Übersetzungen wurden vereinheitlicht.

[10] Qualitätsprüfung der Zielsätze (EN) Nachdem die Übersetzungsfehler außerhalb der KS-Stelle korrigiert wurden, war es erforderlich sicherzustellen, dass die Inhalts- und Stilqualität außerhalb der KS-Stelle (trotz etwaiger Fehler innerhalb der KS-Stelle) akzeptabel sind. Diese Qualitätsprüfung fand über zwei Phasen statt: In der ersten Phase wurde ein Testlauf für alle MÜ-Sätze, die im vorherigen Schritt aufbereitet wurden, durchgeführt. In diesem Testlauf bewertete eine qualifizierte irische Übersetzerin mit sieben Jahren Berufserfahrung die

Tabelle 4.7: Beispiel 5

nach KS	Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, **den Fleck** rückstandslos zu entfernen.
HMÜ Systran Fehlerannotation	*Ever in former times* a mark is treated, *all the more largely* is the probability to remove the mark residueless.
HMÜ Systran Humanevaluation	*The earlier* a mark is treated, *the higher the possibility* of removing the mark without leaving any residue.
RBMÜ Lucy Fehlerannotation	*Each formerly* a spot is treated, *the larger the probability* is to remove the spot *free of residues*.
RBMÜ Lucy Humanevaluation	*The earlier* a spot is treated, *the higher the possibility* of removing the spot *without leaving any residue*.

Die KS-Stelle ist farblich dargestellt: Blau wird für die korrekten Teile der Übersetzung verwendet; rot für die falschen Teile. Fehler *außerhalb* der KS-Stelle, die korrigiert wurden, sind *kursiv und unterstrichen* dargestellt; bzw.***fett, kursiv und unterstrichen*** bei Wortstellungsfehlern *außerhalb* der KS-Stelle, die nicht korrigiert wurden.

Inhalts- und Stilqualität der MÜ-Sätze.[40]

Das Gerät nur an eine vorschriftsmäßig installierte Steckdose anschließen.
Only connect the device to a correctly installed plug socket.

	very low	low	acceptable	high	very high
Content quality	[ ]	[ ]	[ ]	[ ]	[ ]
Style quality	[ ]	[ ]	[ ]	[ ]	[ ]

Comment:

Content quality: The extent to which the translation reflects the information in the source text accurately and in an easy way to understand.

Style quality: The extent to which the translation flows well and sounds natural and idiomatic in Standard Written English.

Abbildung 4.2: Erster Schritt der Qualitätsprüfung

[40] Im Testlauf wurden die Qualitätsdefinitionen vereinfachter bzw. kürzer (als unter §4.5.5.1) angegeben, da es im Test primär um die Darstellung der Ausgangssätze ging.

Wie Abbildung 4.2 zeigt, bestand die Aufgabe im Testlauf darin, die Inhalts- und Stilqualität auf zwei 5-Likert-Skalen zu bewerten und etwaige Fehler bzw. stilistisch kritische Stellen zu kommentieren. MÜ-Sätze, die aufgrund einer Stelle außerhalb der KS-Stelle schlecht bewertet wurden, wurden herausgefiltert. Zwei wesentliche Ergebnisse lieferte dieser Testlauf: Erstens wurde dadurch sichergestellt, dass die MÜ-Sätze *außerhalb* der KS-Stelle keine orthografischen, lexikalischen, grammatischen bzw. semantischen Fehler beinhalten. Zweitens zeigten die Kommentare, dass die bemängelten Stellen primär aus stilistischen Gründen kritisiert wurden.

Auf Basis dieser Ergebnisse wurden die aufbereiteten Zielsätze in der zweiten Phase erneut in einer anderen Testform von zwei Beurteilern geprüft. Der erste Beurteiler ist ein Englischmuttersprachler aus den USA, ein qualifizierter Übersetzer mit mehr als 20 Jahren Berufserfahrung. Die zweite Beurteilerin ist ebenfalls eine Englischmuttersprachlerin, aus Großbritannien, eine qualifizierte Übersetzerin und CELTA-geprüfte Trainerin für Englisch als Fremdsprache. Das Augenmerk lag hierbei auf dem Stil der korrigierten Zielsätze unabhängig von möglichen Fehlern *innerhalb* der KS-Stelle. Mit anderen Worten, die Sätze wurden überprüft, um sicherzustellen, dass die MÜ (außerhalb der KS-Stelle) nach der Fehlerkorrektur mit der minimalen Anzahl von Edits stilistisch akzeptabel blieb. Hierfür wurde die KS-Stelle in jedem Satz markiert und die Bewerter mussten bei jedem Satz ankreuzen, ob die MÜ außerhalb der KS-Stelle stilistisch akzeptabel ist oder nicht (Ja/Nein-Antworten). MÜ-Sätze, die beide Beurteiler als stilistisch inakzeptabel einstuften, wurden von der Humanevaluation ausgeschlossen. MÜ-Sätze, die ein Beurteiler in dieser Phase sowie der Prüfer im Testlauf (d. h. in der ersten Phase) stilistisch kritisierten, wurden ebenfalls von der Humanevaluation ausgeschlossen.

4.5.3.2 Anzahl und Ausgewogenheit der Zielsätze

Nachdem die Zielsätze in den letzten Schritten mehrmals gefiltert wurden, war es erforderlich sicherzustellen, dass die Annotationsgruppen sowohl in der Fehlerannotation als auch in der Humanevaluation in vergleichbaren Verhältnissen repräsentiert sind; beispielsweise wenn bei einer Regel in der Fehlerannotation 30 % der Übersetzung RR, 20 % FF, 25 % FR und 25 % RF waren, wurde bei der Humanevaluation darauf geachtet, dass die vier Gruppen in sehr ähnlichen Verhältnissen repräsentiert sind. Eine vergleichbare Repräsentation der Annotationsgruppen in den beiden Analysen ist essentiell, um einer Verzerrung des Gesamtergebnisses vorzubeugen. Dadurch, dass die Auswahl- bzw. Ausschlusskriterien meistens fehlerorientiert waren, wurden mehr Übersetzungen der Grup-

pen, die Fehler beinhalten, ausgeschlossen, nämlich FF, RF und FR. Aus diesem Grund war die Gruppe RR in der Humanevaluation überpräsentiert und musste reduziert werden. Diese Reduzierung erfolgte gemäß den folgenden Kriterien:

- *Im Falle der Ausgangssätze, die von 3 Systemen oder weniger vor und nach dem Einsatz der KS-Regel richtig übersetzt wurden (Gruppe RR),* wurde der Output eines Systems ausgeschlossen.
- *Im Falle der Ausgangssätze, die von 4 oder 5 Systemen vor und nach dem Einsatz der KS-Regel richtig übersetzt wurden (Gruppe RR),* wurden die Outputs zweier Systeme ausgeschlossen.

Insbesondere bei der Annotationsgruppe RR war der MÜ-Output (meist innerhalb der KS-Stelle) von mehreren Systemen oft identisch. Bei einer Überrepräsentation waren daher diese identischen MÜ-Outputs geeignete Kandidaten für den Ausschluss nach den genannten Kriterien.

Auf Systemebene erfolgte der Ausschluss auf randomisierter Basis, so dass alle Systeme in ungefähr vergleichbaren Prozentsätzen vertreten blieben. Sollte innerhalb einer Regel nach den genannten Kriterien, z. B. bei vier Ausgangssätzen, jeweils der Output eines Systems ausgeschlossen werden, so wurde bei jedem Ausgangssatz randomisiert ein anderes System ausgeschlossen. Abbildung 4.3 zeigt die Aufteilung der Zielsätze in der Humanevaluation bei den einzelnen MÜ-Systemen.

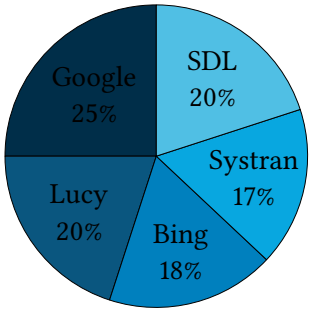


Abbildung 4.3: Aufteilung der Zielsätze in der Humanevaluation auf Systemebene

Tabelle 4.8 zeigt die Verhältnisse der Annotationsgruppen in der Fehlerannotation sowie der Humanevaluation auf Regelebene. Daraus wird ersichtlich, dass die Annotationsgruppen in der Humanevaluation in vergleichbarem Verhältnis wie in der Fehlerannotation stehen.

Tabelle 4.8: Verhältnisse der Annotationsgruppen in der Fehlerannotation sowie der Humanevaluation auf Regelebene

	FR	FF	RF	RR
Regel „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden"				
Fehlerannotation	27,8 %	20,6 %	13,4 %	38,1 %
Humanevaluation	27,5 %	20,8 %	12,5 %	39,2 %
Regel „Partizipialkonstruktion vermeiden"				
Fehlerannotation	9,2 %	37,8 %	23,5 %	29,6 %
Humanevaluation	9,2 %	41,7 %	21,7 %	27,5 %
Regel „Eindeutige pronominale Bezüge verwenden"				
Fehlerannotation	16,9 %	13,0 %	13,0 %	57,1 %
Humanevaluation	17,5 %	13,3 %	11,7 %	57,5 %
Regel „Konditionalsätze mit ‚Wenn' einleiten"				
Fehlerannotation	27,4 %	16,7 %	3,6 %	52,4 %
Humanevaluation	24,2 %	20,8 %	5,0 %	50,0 %
Regel „Keine Wortteile weglassen"				
Fehlerannotation	18,2 %	28,4 %	13,6 %	39,8 %
Humanevaluation	18,3 %	28,3 %	10,8 %	42,5 %
Regel „Passiv vermeiden"				
Fehlerannotation	7,2 %	30,1 %	14,5 %	48,2 %
Humanevaluation	8,3 %	29,2 %	13,3 %	49,2 %
Regel „Funktionsverbgefüge vermeiden"				
Fehlerannotation	29,4 %	20,0 %	8,2 %	42,4 %
Humanevaluation	29,2 %	20,8 %	8,3 %	41,7 %
Regel „Überflüssige Präfixe vermeiden"				
Fehlerannotation	17,4 %	15,2 %	2,2 %	65,2 %
Humanevaluation	13,3 %	15,8 %	3,3 %	67,5 %
Regel „Für zitierte Oberflächentexte gerade Anführungszeichen "..." verwenden"				
Fehlerannotation	32,4 %	40,5 %	1,4 %	25,7 %
Humanevaluation	30,8 %	41,7 %	1,7 %	25,8 %

Die Gesamtzahl der von den Teilnehmern evaluierten MÜ-Sätze betrug 1.100 Sätze. Darüber hinaus gab es insgesamt 545 MÜ-Sätze, die bei mehreren MÜ-Systemen identisch waren (d. h. die Ausgangssätze wurden von verschiedenen Systemen identisch übersetzt). Für jeden Ausgangssatz wurde nur eine Instanz der identischen MÜ-Sätze in der Humanevaluation bewertet, somit waren 95 der 545 Instanzen in den 1.100 humanevaluierten MÜ-Sätzen enthalten. Alle weiteren identischen wiederholten Instanzen (450 von 545) erhielten denselben Score der bewerteten Instanzen. Daher basieren die gelieferten Ergebnisse der Humanevaluation auf einer Gesamtzahl von 1.550 MÜ-Sätzen (1.100 + 450). Tabelle 4.9 präsentiert den Anteil der humanevaluierten MÜ-Sätze von den 240 fehlerannotierten MÜ-Sätzen pro Regel.

Tabelle 4.9: Anteil der humanevaluierten MÜ-Sätze von den 240 fehlerannotierten MÜ-Sätzen pro Regel

Für zitierte Oberflächent. gerade Anführungsz. verwenden	62 %
Funktionsverbg. vermeiden	70 %
Konditionals. mit ‚Wenn' einleiten	70 %
Eindeutige pronomin. Bezüge verwenden	64 %
Partizipialkonst. vermeiden	82 %
Passiv vermeiden	69 %
Konstr. mit „sein + zu + Infinitiv" vermeiden	81 %
Überflüssige Präfixe vermeiden	77 %
Keine Wortteile weglassen	73 %

N der Humanevaluation = 1.550 MÜ-Sätze bestehend aus 1.100 humanevaluierten MÜ-Sätzen plus 450 wiederholten MÜ-Sätzen

Wie Tabelle 4.9 zeigt, wurden bei den Regeln über 60 % der annotierten MÜ-Sätze humanevaluiert. Der Anteil der evaluierten MÜ-Sätze bewegte sich zwischen 62 % und 82 %.

4.5.4 Design der Fehlerannotation

Das Ziel der Fehlerannotation liegt darin, die Übersetzungsfehler vor und nach dem Einsatz der einzelnen KS-Regeln zu identifizieren und zu vergleichen. Die Fehlerannotation lieferte Ergebnisse zu Fehleranzahl und -typen sowie zur Erscheinung bzw. Eliminierung von bestimmten Fehlertypen in Zusammenhang mit einer bestimmten Regel. Darüber hinaus wurden die Ergebnisse der Fehlerannotation mit denen der Humanevaluation trianguliert (siehe §4.5.5.4).

Wie unter §4.5.3.1 detailliert erklärt, umfasst der Datensatz zwei Versionen von jedem Ausgangssatz: Die *vor-KS-Version* stellt den Ausgangssatz mit dem Verstoß gegen eine der neun analysierten KS-Regeln dar. In der *nach-KS-Version* wurde der Verstoß durch den Einsatz der entsprechenden KS-Regel behoben.[41] Bei der Fehlerannotation wurden die Fehler in der maschinellen Übersetzung beider Versionen identifiziert und einem Fehlertyp zugeordnet. Die Auswahl und Festlegung der Fehlertaxonomie sowie der Ablauf der Fehlerannotation werden im Folgenden genauer erläutert.

4.5.4.1 Auswahl und Festlegung der Fehlertaxonomie

Vorherige Studien, wie in der Diskussion der Fehlerannotationsmethode (unter §3.4.3.1) detailliert dargestellt, bieten zahlreiche Fehlertaxonomien an, die sich jedoch im Kern ähneln. Dadurch, dass die Taxonomien für unterschiedliche Zwecke und Zielgruppen entwickelt wurden, variieren sie in Hinsicht auf ihren Granularitätsgrad. Einige Fehlertaxonomien sind detailliert (vgl. Flanagan 1994; Correa 2003), andere wenden wenige aber breite und mehrstufige Klassifikationen (vgl. Vilar u. a. 2006; Farrús u. a. 2010) an. Infolge dieser Diversität zielten Vilar u. a. (2006) mit ihrer Fehlertaxonomie darauf ab, eine explizite Fehlerklassifikation zu entwickeln. Vilar et al. (ebd.) lieferten eine dreistufige Fehlertaxonomie, die aus fünf Hauptkategorien (fehlende Wörter, Wortstellungsfehler, falsche Wörter, unbekannte Wörter und Zeichensetzungsfehler) besteht und häufig in der Evaluationsforschung verwendet wird (u.a. in Avramidis & Koehn 2008, Bojar 2011 und Popović & Burchardt 2011). Aufgrund der Explizität, Übersichtlichkeit sowie des angemessenen Granularitätsgrads und Umfangs der Fehlertaxonomie von Vilar u. a. (2006) wurde sie in der vorliegenden Studie als Basis der Fehlerannotation angewendet. Der Umfang der in dieser Taxonomie enthaltenen Fehlertypen war für das Studienziel angebracht, da es in der Studie nicht allgemein um die Bewertung von MÜ-Outputs verschiedener Systeme geht, sondern um eine explizite Bewertung der Fehler, die in Zusammenhang mit den analysierten KS-Regeln auftreten. Die Anwendung der Fehlertaxonomie von Vilar et al. (ebd.) schließt jedoch nicht aus, dass eine andere umfangreichere Taxonomie, wie z. B. das MQM-Framework, zur Analyse herangezogen werden kann. Dies wäre insbesondere bei der Analyse von weiteren KS-Regeln, die feinkörnige bzw. weitere spezifische Fehlertypen erfordern, sinnvoll.

Die Festlegung einer Fehlertaxonomie fand auf Basis eines Bottom-up-Ansatzes in zwei Schritten statt. Im ersten Schritt wurde die Fehlertaxonomie von Vilar

[41] Für mehr zu der Vorgehensweise beim Einsatz der Regeln siehe §4.5.3.1 (Schritt [5]) sowie §4.5.2.2, die die Darstellung der analysierten Regel genauer erläutert.

u. a. (2006) zur Orientierung herangezogen. In diesem Schritt wurden bei jeder Regel sieben Sätze annotiert. Auf Basis dieser Annotation stellte es sich heraus, dass die Fehlertaxonomie von Vilar et al. einen Fehlertyp „Redewendungen" (idiomatic expressions)[42] beinhaltet, der in den analysierten MÜ-Sätzen nicht auftrat. Des Weiteren führen Vilar et al. einen Fehlertyp für den „Stil" auf. In dieser Studie wird der Stil nicht als Fehlertyp betrachtet, sondern als ein Übersetzungsproblem, dessen Einfluss im Rahmen der Humanevaluation gesondert bewertet wird. Beide Fehlertypen (Redewendungen und Stil) werden von Vilar et al. (ebd.: 698) als „less important" bezeichnet, dennoch wurden sie nicht wegen ihrer Bedeutung, sondern aus den genannten Gründen ausgeschlossen. Auf der anderen Seite kamen drei Fehlertypen vor, die nicht direkt in der Fehlertaxonomie von Vilar et at. jedoch im Datensatz vertreten sind. Diese Fehler sind „Großschreibung", „Konsistenzfehler" und „Kollokationsfehler". Die drei Fehler traten überwiegend bei drei Regeln[43] auf, daher war es erforderlich, sie explizit hinzuzufügen, damit sie direkt (d. h. nicht als Unterkategorie) analysiert werden. Tabelle 4.10 listet die Fehlertypen nach Vilar et al. gegenüber den aus dem ersten Analyseschritt resultierenden Fehlertypen auf.

In der dreistufigen Fehlertaxonomie von Vilar u. a. (2006) wurden zum Teil die Fehlertypen und ihre Ursachen gemischt. Die erste Stufe beinhaltet z. B. die Kategorien „Missing words" und „Unknown words". In der zweiten Stufe folgen bei „Missing words" die Unterkategorien „Content words" und „Filler words" (d. h. eine genauere Spezifizierung des Fehlertyps), während die „Unknown words" in der zweiten Stufe in „Unknown stem" und „Unseen forms" (Ursache des Fehlers) unterteilt werden. Der Einsatz einer mehrstufigen Fehlertaxonomie hätte die Analyse unübersichtlich und unnötig kompliziert gestaltet. Daher wurden die identifizierten Fehlertypen im zweiten Schritt auf einer Stufe gruppiert und den relevanten linguistischen Hauptkategorien (Orthografie, Lexik, Grammatik, Semantik) zugeordnet. Diese linguistisch motivierte Klassifizierung war dem Studienziel dienlich, denn auf diese Weise konnte bei jeder KS-Regel genauer analysiert werden, in welchem linguistischen Zweig die Fehler ihre Wurzel haben.

Nach der oben geschilderten Vorgehensweise wurde die angewandte Fehlertaxonomie festgelegt, mit der die Übersetzungsfehler vor und nach dem Einsatz der KS-Regel identifiziert und verglichen werden können. Dadurch, dass die Taxonomie auf Basis des analysierten Datensatzes und Sprachenpaars festgelegt wurde, kann eine Anpassung bei der Analyse anderer Texttypen bzw. Sprachenpaare

[42] Das Beispiel von Vilar u. a. (2006) dafür war: „It's raining cats and dogs."

[43] Großschreibungsfehler bei der Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden"; Konsistenzfehler bei „Eindeutige pronominale Bezüge verwenden" und Kollokationsfehler bei „Funktionsverbgefüge vermeiden".

Tabelle 4.10: Fehlertypen nach Vilar u. a. (2006) gegenüber den der vorliegenden Studie

Fehlertypen	**Vilar u. a. (2006)** (Bezeichnung nach Vilar)	**Vorliegende Studie**
Orthografie		
Zeichensetzungsfehler	Ja (Punctuation)	Ja
Großschreibungsfehler	Nein	Ja
Lexik		
Wort ausgelassen	Ja (Missing words)	Ja
Wort extra eingefügt	Ja (Extra words)	Ja
Wort unübersetzt geblieben	Ja (Unknown words)	Ja
Konsistenzfehler	Nein	Ja
Grammatik		
falsche Wortart / Wortklasse	Ja (Incorrect form)	Ja
falsches Verb	Ja (Incorrect form)	Ja
Kongruenzfehler	Ja (Incorrect form)	Ja
falsche Wortstellung	Ja (Word order)	Ja
Semantik		
Verwechselung des Sinns	Ja (Sense)	Ja
falsche Wahl	Ja (Wrong choice)	Ja
Kollokationsfehler	Nein	Ja
Redewendungen	Ja (Idioms)	Nein
Stil	Ja (Style)	Nein

(vgl. ebd.: 701) erforderlich sein. Insgesamt besteht die angewandte Fehlertaxonomie aus den folgenden 13 Fehlertypen (Tabelle 4.11).

Auf *orthografischer* Ebene wurden *Zeichensetzungs-* sowie *Großschreibungsfehler* identifiziert. Bei Sätzen, die mehrere Zeichensetzungsfehler (z. B. zwei fehlende Kommas) beinhalten, wurde jeder Fehler als *ein* Fehler berechnet. *Zeichensetzungsfehler* können bei allen Regeln auftreten, denn die beiden analysierten Sprachen verfügen über unterschiedliche Zeichensetzungsregelungen. Die *Großschreibung* war insbesondere bei der Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ von Bedeutung, denn zitierte Oberflächentexte sollten großgeschrieben werden.

Auf *lexikalischer* Ebene werden die drei möglichen lexikalischen Fehler abge-

Tabelle 4.11: Die analysierten Fehlertypen

	Orthografie
1	Zeichensetzungsfehler
2	Großschreibungsfehler
	Lexik
3	Wort ausgelassen
4	Wort extra eingefügt
5	Wort unübersetzt geblieben
6	Konsistenz
	Grammatik
7	falsche Wortart/Wortklasse
8	falsches Verb
9	Kongruenzfehler
10	falsche Wortstellung
	Semantik
11	Verwechslung des Sinns
12	falsche Wahl
13	Kollokationsfehler

deckt: Ein Wort im Ausgangssatz fehlte in der Übersetzung (*Wort ausgelassen*); ein Wort wurde zusätzlich übersetzt, obwohl es nicht im Ausgangssatz enthalten war (*Wort extra eingefügt*); ein Wort wurde vom System in der Ausgangssprache wiedergegeben (*Wort unübersetzt geblieben*). Letzteres wird in der MÜ z. B. zur 1:1-Wiedergabe von Eigennamen oder innovativen fremdsprachlichen Termini eingesetzt. Sollte jedoch ein Wort unübersetzt bleiben, obwohl eine Übersetzung dafür erforderlich und möglich ist, zählt dieser Fall als Fehler. Ferner wurde der Fehlertyp *Konsistenzfehler* miteinbezogen, da er in den analysierten Daten (insbesondere bei der Regel „Eindeutige pronominale Bezüge verwenden“) vorkam. Ein lexikalischer Konsistenzfehler tritt auf, wenn dasselbe Wort im Satz unterschiedlich übersetzt wird (vgl. Mertin 2006: 249). Ein Beispiel hierfür ist die Übersetzung des Worts „Gerät“ im Hauptsatz als ‚device‘ und im Nebensatz als ‚appliance‘. Eine mögliche Ursache für diesen MÜ-Fehler ist die Verwendung von Synonymen, um eine Wiederholung zu vermeiden und somit den Stil zu verbessern. Dennoch hat die Konsistenz im Falle von technischen Texten oberste Priorität, da sie die Lesbarkeit und Verständlichkeit unerlässlich fördert.

Auf *grammatischer* Ebene wurden vier Fehlertypen definiert: Die ersten drei Fehlertypen decken die Fälle ab, in denen das MÜ-System falsche Wortformen liefert: erstens auf *Wortartebene*, indem z. B. die MÜ ein Adverb anstelle eines Adjektivs enthält. Zweitens eine falsche Übersetzung von *Verben* in Zeitform oder Konstruktion. Dieser Fehlertyp trat zwar bei mehreren KS-Regeln auf, jedoch war er besonders relevant bei der Regel „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden", da sich die Übersetzung der Passiv-Ersatzkonstruktion oft als problematisch erwies. Im dritten Fall geht es um den *Kongruenzfehler*. Hierbei wurde nach Duden[44] die „Übereinstimmung zusammengehörender Teile im Satz (in Kasus, Numerus, Genus und Person)" sowie die „inhaltlich sinnvolle Vereinbarkeit des Verbs mit anderen Satzgliedern" geprüft. Der letzte grammatische Fehlertyp ist die *falsche Wortstellung*. Ein Wortstellungsfehler kann in verschiedenen Fällen vorkommen, daher ist er für alle KS-Regeln relevant. Bei der Annotation wurde hierbei berücksichtigt, dass der Wortstellungsfehler auf Basis der Mindestzahl an erforderlichen Bewegungen bzw. Anpassungen zur Behebung der Fehler berechnet wird. Sollte der Fehler ein Bestandteil der KS-Stelle sein, bedeutet das, dass weitere Satzteile außerhalb der KS-Stelle ebenfalls falsch positioniert sind. In einem solchen Fall wurde jeweils ein Fehler innerhalb als auch außerhalb der KS-Stelle berechnet.

Auf *semantischer* Ebene wurden drei Fehlertypen annotiert: Erstens die *Verwechslung des Sinns*. Bei diesem Fehlertyp liefert das MÜ-System eine der möglichen Übersetzungen des Worts, jedoch gibt diese Übersetzung kontextuell nicht die zutreffende Bedeutung wieder (z. B. die Übersetzung des Verbs ‚belegen' (in ‚Belegen Sie das Kaufdatum durch eine Kaufquittung') als ‚occupy' anstelle von ‚verify' oder ‚prove'. Der zweite semantische Fehlertyp ist eine *falsche Wahl*. Dieser Fehlertyp tritt auf, wenn die gelieferte MÜ keine mögliche Übersetzung für das betroffene Wort ist (z. B. ‚die Firmware-Version' wurde als ‚the firmware design' übersetzt). Der letzte Fehlertyp ist der Kollokationsfehler. Duden[45] definiert die Kollokation als die „inhaltliche Kombinierbarkeit sprachlicher Einheiten" sowie der „Zusammenfall verschiedener Inhalte in einer lexikalischen Einheit". Nach dieser Definition wurde in den MÜ-Sätzen geprüft, ob vorhandene deutsche Kollokationen ins Englische richtig übersetzt wurden. Dieser Fehlertyp war insbesondere für die KS-Regel „Funktionsverbgefüge vermeiden" von Bedeutung.

[44] Online unter: https://www.duden.de/rechtschreibung/Kongruenz
[45] Online unter: https://www.duden.de/rechtschreibung/Kollokation

4.5.4.2 Beschreibung und Ablauf bei der Annotation

Die Fehlerannotation wurde manuell mithilfe der oben beschriebenen Fehlertaxonomie über zwei Phasen durchgeführt.[46] In der ersten Phase wurden die 2.160 MÜ-Sätze von einem DE-EN-Übersetzer mit sechs Jahren Berufserfahrung annotiert. In der zweiten Phase wurde die Annotation des ersten Übersetzers von zwei weiteren erfahrenen DE-EN-Übersetzern (10+ Jahre Berufserfahrung) geprüft. Aufgrund der großen Anzahl der Sätze prüfte jeder Übersetzer (in der zweiten Phase) die Hälfte der Sätze. Die Prüfung fand in Form einer Auswahlaufgabe statt, in der jeder Bewerter ankreuzen musste, ob er der Annotation (in Hinsicht auf die Identifikation eines Fehlers sowie den zugeordneten Fehlertyp) zustimmt oder nicht. Im Fall dass ein Bewerter der Annotation eines MÜ-Satzes nicht zustimmte, musste er die MÜ erneut annotieren. Die Prozentsätze der erneut annotierten MÜ waren 27 % beim ersten Bewerter bzw. 31 % beim zweiten. Anschließend musste der andere Bewerter die neue Annotation prüfen und sich für eine der beiden Annotationen (Annotation aus der ersten Phase oder Annotation des anderen Bewerters in der zweiten Phase) entscheiden.

Der Annotator bzw. die Bewerter sollten sich die Mindestzahl an Anpassungen bzw. Bewegungen vorstellen, die zur Korrektur der MÜ-Fehler erforderlich ist und eine Veröffentlichung des Texts ermöglicht. Während der beiden Phasen wurde die Ausgangssätze zur Verfügung gestellt. Andernfalls wäre die Annotation von mehreren Fehlertypen nicht möglich gewesen. Bei der Annotation zählte jeder Fehler als *ein* Fehler. Sollte ein Wort mehrere Fehlertypen aufweisen, wurden alle Fehlertypen berücksichtigt. Beispielsweise kann ein Wort einen Großschreibungsfehler sowie einen Wortstellungsfehler aufweisen. In diesem Fall wurden zwei Fehler berechnet. Je nachdem an welcher Stelle der Fehler auftrat, wurde bei der Analyse zwischen Fehlern innerhalb der KS-Stelle und Fehlern außerhalb der KS-Stelle unterschieden.[47]

4.5.4.3 Struktur der Ergebnisse der Fehlerannotation

Im Rahmen der Fehlerannotation erfolgte der Vergleich der Szenarien vor-KS vs. nach-KS im Hinblick auf die Fehleranzahl und den Fehlertyp sowie dichotom

[46] Aufgrund des speziellen Aufbaus der Fehlerannotation, insbesondere bei der Unterscheidung zwischen Fehlern innerhalb und außerhalb der KS-Stelle, konnte die Annotation nicht mit einem klassischen Annotationstool durchgeführt werden. Daher wurde die Annotation mithilfe von Microsoft Excel durchgeführt. In Excel wurden Formeln für die Berechnung der Fehler hinterlegt und sämtliche quantitativen Daten errechnet.

[47] Die KS-Stelle ist der Teil des Ausgangssatzes, der bei dem Einsatz der KS-Regel modifiziert werden muss, und sein Äquivalent im Zielsatz, genauer beschrieben unter §4.5.2.1.

im Sinne von Existenz/Vorhandensein oder Nichtexistenz/Nichtvorhandensein der Fehler. Letzteres wurde bei der Bildung von vier sog. „Annotationsgruppen" benutzt. Konkret lieferte die Fehlerannotation – unter Berücksichtigung der Aufteilung „Fehler innerhalb der KS-Stelle" und „Fehler außerhalb der KS-Stelle" – die folgenden Ergebnisse:

(1) Die Fehleranzahl vor und nach der Anwendung der einzelnen KS-Regeln. Hierfür wurde der Signifikanztest Wilcoxon verwendet, da die Variablen ordinal sind.

(2) Die Fehlertypen, die vor und nach der Anwendung der einzelnen KS-Regeln, auftraten bzw. eliminiert wurden. Für den Vergleich der Fehlertypen wurde der Signifikanztest McNemar verwendet, da er ermöglicht, zwei verbundene dichotome Parameter zu vergleichen; somit konnten signifikante Veränderungen bei jedem Fehlertyp vor- vs. nach-KS identifiziert werden.

(3) Eine Kategorisierung der Ergebnisse in vier Annotationsgruppen FR, RF, RR und FF:

   (a) MÜ der KS-Stelle ist vor der Anwendung der KS-Regel falsch und nachher richtig (FR).

   (b) MÜ der KS-Stelle ist vor der Anwendung der KS-Regel richtig und nachher falsch (RF).

   (c) MÜ der KS-Stelle ist sowohl vor als auch nach der Anwendung der KS-Regel richtig (RR).

   (d) MÜ der KS-Stelle ist sowohl vor als auch nach der Anwendung der KS-Regel falsch (FF).

Die Häufigkeiten der Annotationsgruppen wurden mit Bootstrapping[48] berechnet.

(4) Eine Untersuchung der Fehleranzahl außerhalb der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu davor bei den RR-Fällen. In anderen Worten sollte die KS-Stelle sowohl vor als auch nach dem Einsatz der KS-Regel fehlerfrei sein, wird der MÜ-Output außerhalb der KS-Stelle in Hinsicht auf die Veränderung in der Fehleranzahl genauer untersucht. Die

[48] Bootstrapping ist ein statistisches Verfahren zur Schätzung der Stichprobenverteilung eines Schätzers durch erneute Stichprobenerstellung mit Ersatz aus der ursprünglichen Stichprobe. Es wird als ein nützliches Verfahren zum Testen der Modellstabilität betrachtet. (IBM o.D.)

Untersuchung wurde auf Sprachenpaarebene und MÜ-Systemebene (siehe §5.3.3 bzw. §5.5.3) durchgeführt. Hierfür wurde ebenfalls der Signifikanztest Wilcoxon verwendet.

4.5.5 Design der Humanevaluation

Die Fehlerannotation lieferte objektive quantitative Daten über die aufgetretenen Fehler, ihre Anzahl und ihren Typ vor und nach dem Einsatz der KS-Regeln. Dennoch bleiben die Fragen offen, ob ein Rückgang der Fehleranzahl zwangsläufig auch auf eine bessere MÜ hindeutet oder ob bestimmte Fehlertypen die Qualität der MÜ verhältnismäßig stärker beeinflussen und wie genau die Qualität beeinflusst wird. Eine quantitative Analyse der Ergebnisse der Fehlerannotation kann alleine keinen Aufschluss über diese Details geben, um das Bild vor-KS versus nach-KS in Bezug auf die Qualität vergleichen zu können. Hierfür ist der Forscher auf subjektive qualitative Bewertungen angewiesen. Daher war die Anwendung einer Humanevaluation erforderlich.

Ziel der Humanevaluation ist der Vergleich der Qualität des Inhalts und des Stils der KS-Stelle (nicht der Qualität des ganzen Satzes) vor versus nach dem Einsatz der einzelnen KS-Regeln.[49]

Obwohl eine gewisse Korrelation zwischen der Stilqualität und der Inhaltsqualität zu erwarten ist (z. B. ist ein Satz, der einen grammatischen Fehler beinhaltet bzw. unverständlich ist, i. d. R. nicht idiomatisch), wurden die Stilqualität und der Inhaltsqualität getrennt bewertet. Der Grund für diese separate Bewertung liegt darin, dass die einzelnen KS-Regeln die Stilqualität und Inhaltsqualität unterschiedlich beeinflussen können. Bei einigen Regeln (wie z. B. „Funktionsverbgefüge vermeiden“ und „Partizipialkonstruktionen vermeiden“) kann der Stil eine wesentliche Rolle bei der Qualitätsbewertung spielen. In Szenarien, in denen die Übersetzungen sowohl vor als auch nach dem Einsatz der KS-Regel korrekt sind, kann die Stilqualität variieren. Dementsprechend ist eine Differenzierung bei der Bewertung des Einflusses zentral für die Ergebnisse.

Die Triangulation der Ergebnisse der Fehlerannotation mit denen der Humanevaluation ermöglichte die Untersuchung der Korrelation zwischen den Fehlertypen und der Qualität sowie die Untersuchung der Qualität der Annotationsgruppen. Folglich konnten durch die triangulierten Ergebnisse die Qualitätsver-

[49] Die Vergleichswerte der Inhalts- und Stilqualität der KS-Stellen in den vor-KS- und nach-KS-Szenarien konnten ermittelt werden, indem der MÜ-Output vor und nach dem Einsatz der KS-Regeln mit Ausnahme der KS-Stelle vereinheitlicht und die Differenz zwischen den Qualitätsbewertungen (Scores) vor und nach KS berechnet wurde (die genaue Aufbereitung der Zielsätze für die Humanevaluation ist unter §4.5.3.1 (Schritt [9]) dargestellt).

änderung und somit der KS-Einfluss genauer beleuchtet werden. Im Folgenden wird das Testdesign genauer erläutert.

4.5.5.1 Definition der Qualität

Wie in der Literaturübersicht unter §3.4.1 dargestellt, wenden die MÜ-Evaluationsstudien zahlreiche Synonyme von Qualitätskriterien bzw. sich überlappenden Kriterien an. Das häufigste Qualitätskriterium in den geschilderten Studien ist die Verständlichkeit.

Wie Lehrndorfer (1996a: 339) es beschreibt, ist die Verständlichkeit „[e]in unerschöpfliches Thema in den Diskussionen über technische Dokumentation". Dabei werden Aspekte (sog. „Verständlichkeitsmacher") untersucht, die das Verstehen des technischen Inhalts erleichtern (ebd.). Zu den bekanntesten Verständlichkeitsmodellen für die Fachkommunikation zählen: das Hamburger Modell von Langer u. a. (1974), das vier Merkmalsdimensionen der Verständlichkeit (Einfachheit, Gliederung/Ordnung, Kürze/Prägnanz und zusätzliche Stimulanz) definiert. Dieses Modell wurde jedoch aufgrund seiner textzentrierten Perspektive (d. h. Leser wird außer Acht gelassen) kritisiert und daraufhin von Groeben (1982) und Göpferich (2001) erweitert. Nach dem interaktionalen Ansatz von Groeben (1982) ist die Textverständlichkeit eine Interaktion zwischen Text und Leser. Die vier Dimensionen der Textverständlichkeit von Groeben (sprachliche Einfachheit, semantische Kürze/Redundanz, kognitive Gliederung/Ordnung und stimulierender kognitiver Konflikt) stimmen weitgehend mit denen des Hamburger Modells überein und erweitern es zugleich. In dem Karlsruher Verständlichkeitskonzept erweiterte Göpferich (2001) beide Modelle um die Kommunikationssituation (kommunikative Funktion des Texts) und entwickelte damit einen „kommunikationsorientiert-integrativen Ansatz zur Bewertung der Verständlichkeit von Texten" bestehend aus sechs Dimensionen (Struktur, Simplizität, Motivation, Prägnanz, Korrektheit und Perzipierbarkeit). Es ist hier aus Platzgründen nicht möglich, den zahlreichen Modellen und Konzepten der Verständlichkeit gerecht zu werden und diese tiefergehend zu diskutieren. Die Arbeit beschränkt sich daher auf die computerlinguistische Perspektive der Verständlichkeit, wie sie in Hutchins & Somers's (1992) bekanntem Werk zur MÜ-Evaluation definiert ist.

Im Gegenteil zur Verständlichkeit wird der Stil nur in wenigen Evaluationen miteinbezogen. Hutchins & Somers (1992: 164) finden bei der MÜ-Qualitätsevaluation jedoch, dass „the appropriateness of a particular style is an important factor". Nicht nur für die MÜ-Qualität im Allgemeinen, sondern auch für die technische Dokumentation im Speziellen spielt der Stil eine wesentliche Rolle

bei der Förderung der Verständlichkeit und der Genauigkeit, wie Püschel (1996: 307ff.) in seinem Beitrag „Sprachstil – ein Thema für Technische Redakteure?“ genauer erklärt. Nach Püschel (1996: 307) gehören zum Stil einerseits, „dass der Schreibstil viele Facetten hat“ und andererseits, „dass man das Gleiche nicht auf unterschiedliche Weise sagen kann“. Zur Illustration dieser Eigenschaften führt Püschel (1996: 317) folgende Beispielsätze aus der technischen Dokumentation ein: [1] „…ist das Gerät unbedingt zu entkalken“; [2] „…muss man das Gerät unbedingt entkalken“; [3] „…müssen Sie das Gerät unbedingt entkalken“. In der Infinitivkonstruktion [1] wird die Person, die die Handlung durchführen soll, nicht ausgedrückt. In der syntaktisch vollständigen Version [2] bleibt die handelnde Person vage; sie wird mit ‚man‘ ausgedrückt. In der letzten Version [3] wird die handelnde Person ausgedrückt bzw. direkt mit ‚Sie‘ angesprochen. Mithilfe dieser Beispiele verdeutlicht Püschel (1996: 317), wie „der Zusammenhang von Gedanke und Stil kein Glasperlenspiel“ ist. Er betrachtet kritisch die Trennung von Verständlichkeit und Stil in der technischen Dokumentation:

> [w]ann immer Probleme des technischen Schreibens diskutiert werden, geht es auch um die Verständlichkeit von Texten und um die Frage, wie Texte verständlicher gemacht werden können. Es liegt dabei auf der Hand, dass jede Optimierungsmaßnahme einen Eingriff in die Gestalt des Texts bildet; […], dass mit den Veränderungen im Stil zwangsläufig inhaltliche Veränderungen einhergehen. (ebd.)

Im Hinblick auf die KS verdeutlicht die Diskussion von Püschel, wie der Stil in der technischen Dokumentation (in den obigen Beispielen bei den Regeln zum Vermeiden des Passivs und des Passiversatzes) eine Rolle spielt und wie der Stil Hand in Hand mit der Verständlichkeit und der Genauigkeit des Inhalts einhergeht und sie entsprechend fördern kann (vgl. Püschel 1996: 335). Vor diesem Hintergrund durfte die Stilqualität bei dieser Studie nicht außer Acht gelassen werden.

Neben der Verständlichkeit (vgl. „intelligibility“ bei White 2003; vgl. „comprehensibility“ und „clarity“ bei Vanni & Miller 2002 sowie King u. a. 2003) und dem Stil (vgl. Vanni & Miller 2002 sowie King u. a. 2003) ist die Genauigkeit das dritte Qualitätskriterium (vgl. „fidelity“ bei White 2003; vgl. „accuracy“ bei Arnold 1994: 169), das in die MÜ-Evaluationen miteinbezogen wird.

Eine Definition der MÜ-Qualität, die die Verständlichkeit, die Genauigkeit und den Stil abdeckt und gleichzeitig die Problematik der Bezeichnungen, Aufteilung und möglichen Überschneidungen der Qualitätskriterien (vgl. §3.4.1) berücksichtigt, ist die von Hutchins & Somers (1992: 163). Unter Angabe der zutreffenden Synonyme definieren sie die drei genannten Qualitätskriterien wie folgt:

> (a) Fidelity or accuracy, the extent to which the translated text contains the ‚same'' information as the original; (b) Intelligibility or clarity, the ease with which a reader can understand the translation; and (c) Style, the extent to which the translation uses the language appropriate to its content and intention. (ebd.)

Damit liefern Hutchins und Somers in sich geschlossene und klar aufgeteilte Qualitätskriterien (vgl. Fiederer & O'Brien 2009). Gleichzeitig ist es nachvollziehbar – wie die obigen Beispiele von Püschel zeigen –, dass ein gewisser gegenseitiger Einfluss sowie eine potenzielle Überschneidung der Kriterien nicht auszuschließen sind.

Eine weitere aktuelle Qualitätsevaluationsmethode, die häufig angewandt wird, ist das MQM (Multidimensional Quality Metrics) Framework (Lommel u. a. 2013). Es handelt sich dabei um ein umfassendes Framework aus den fünf Hauptdimensionen „accuracy, fluency, design, verify und internationalization" und mehr als 100 unterkategorisierten Fehlertypen (mehr dazu unter §3.4.3.1). Ausgehend von der Annahme, dass kein einzelnes festes Schema zur Qualitätsbewertung bei unterschiedlichen Übersetzungsprojekten verwendet werden kann, sollen für jedes Projekt je nach seinen Anforderungen und Erwartungen die relevanten Elemente maßgeschneidert ausgewählt werden. (Lommel u. a. 2013) Diese Methode wäre insbesondere bei der Analyse einer Vielzahl von KS-Regeln, die feinkörnige bzw. weitere spezifische Fehlertypen erfordern, sinnvoll. Bei der vorliegenden Studie waren die Integrität und das Granularitätsniveau die zwei Faktoren, die zur Auswahl der MÜ-Qualitätsdefinition von Hutchins & Somers (1992) führten.

Die Bewertung in der vorliegenden Studie findet auf inhaltlicher und stilistischer Ebene statt. Da der Stil und der Inhalt der MÜ vom Einsatz der einzelnen KS-Regeln zu unterschiedlichem Grad beeinflusst werden können, ist eine Differenzierung bei der Analyse des Einflusses der einzelnen Regeln auf die detaillierten Kriterien der Stil- und Inhaltsqualität essentiell für die Studienergebnisse. Im Folgenden werden die angewandten Definitionen der Stil- und Inhaltsqualität detailliert dargestellt.

Den Teilnehmern der Humanevaluation wurde folgende Definition der Inhaltsqualität präsentiert (vgl. Hutchins & Somers 1992: 163):

> The extent to which the translation reflects the information in the source text accurately; and the extent to which the translation is easy to understand. (ebd.)

Genauigkeit und Verständlichkeit wurden als Kriterien der Inhaltsqualität (Abbildung 4.6 [3b]) getrennt bewertet, damit der Einfluss der KS-Regeln auf die bei-

den Qualitätskriterien nach Hutchins und Somers beurteilt werden kann. Gleichzeitig wurden die Kriterien Genauigkeit und Verständlichkeit auf der Skala der Inhaltsqualität zusammengeführt, denn auf ihrer Grundlage wurde untersucht, inwiefern der *Inhalt* der MÜ originaltreu und verständlich ist.

Anhand des Inhaltsqualität-Scores auf der Likert-Skala wurde die Inhaltsqualität operationalisiert und entsprechend wurde der Einfluss des Einsatzes der KS-Regeln auf die Inhaltsqualität quantitativ bewertet. Um sicherzustellen, dass die Bewertungen der Verständlichkeit bzw. der Genauigkeit in den Score der Inhaltsqualität einfließen, wurden die Bewerter in den Testanweisungen aufgefordert, erst die relevanten Qualitätskriterien (siehe [3b] in Abbildung 4.6) anzukreuzen und ihre Kriterienauswahl zu begründen ([4] in Abbildung 4.6) und danach den Score der Inhaltsqualität auf Basis der ausgewählten Kriterien zu vergeben ([2] in Abbildung 4.6) (siehe „Testaufgaben" unter §4.5.5.2).

Ein positiver Nebeneffekt der Zusammenführung der beiden Kriterien ist die Reduzierung des Zeitaufwands und der Komplexität der Bewertungsaufgabe für die Bewerter. Die Bewerter müssen zwar einen potenziellen Einfluss auf die einzelnen Kriterien erkennen (siehe [3b] in Abbildung 4.6), jedoch nicht den Schwierigkeitsgrad der identifizierten Fehler differenziert einmal in Hinsicht auf die Genauigkeit und getrennt in Bezug auf die Verständlichkeit quantitativ bewerten, was sich aufgrund der großen Anzahl der bewerteten Sätze als vorteilhaft erwies.

Für die Stilqualität wurde den Teilnehmern der Humanevaluation folgende Definition präsentiert (vgl. Hutchins & Somers 1992: 163; Fiederer & O'Brien 2009: 57):

> The extent to which the translation sounds natural and idiomatic in Standard Written English, is appropriate to the intention of its content as well as is presented clearly orthographically. (ebd.)

Die Stilqualitätsdefinition zeigt, wie Püschel (1996: 335) betont, dass der Stil „keine Sprachkosmetik" sei. Die Definition umfasst drei Qualitätskriterien, die die Verständlichkeit und Genauigkeit fördern können (vgl. obige Beispiele von Püschel 1996: 317): die Idiomatik der MÜ, die Eignung der MÜ für die Intention des Inhalts sowie die korrekte bzw. klare orthografische Darstellung der MÜ.[50] Im Folgenden werden die drei Kriterien der Stilqualität näher erläutert:

(1) *Die Idiomatik der MÜ*: Typischerweise wird der MÜ-Output post-editiert. Nach dem klassischen Ablauf wird ein „Light Post-Editing" zur Korrektur

[50] Es existieren zahlreiche Definitionen für die Begriffe Stil und Intention, die aus Platzgründen nicht alle diskutiert werden können. Die Studie beschränkt sich auf die computerlinguistische Perspektive dieser Begriffe, wie in diesem Abschnitt erklärt.

der wesentlichen Fehler durchgeführt mit dem Ziel, den Inhalt in einer verständlichen und genauen Form bereitzustellen (vgl. O'Brien u. a. 2009; O'Brien 2010b). Sollte der Stil optimiert werden, so erfolgt ein „Full Post-Editing", bei dem aus dem MÜ-Output eine stilistisch vergleichbare Humanübersetzung erzeugt werden soll (vgl. Wagner 1987). Das ist bisher der gängige Ablauf, da die früheren Ansätze überwiegend mit schwerwiegenden Fehlern (u. a. im Bereich der Morphologie und Grammatik, wie z. B. Wortstellungsfehlern) zu kämpfen haben. Mit der Entwicklung der NMÜ haben sich die Eigenschaften des MÜ-Outputs wesentlich geändert, denn die NMÜ kann hingegen diese Schwierigkeiten lösen und darüber hinaus eine im Wesentlichen flüssige Übersetzung liefern (vgl. Bentivogli u. a. 2016; Toral & Sanchez-Cartagena 2017). Die hohe Flüssigkeit des NMÜ-Outputs zählt zu den dominanten Stärken dieses Ansatzes (vgl. ebd.). Angesichts dieser Entwicklung darf bei einer entwicklungs- bzw. forschungsstandgemäßen empirischen Untersuchung die Evaluation der Idiomatik nicht fehlen. Anderenfalls könnte die technische Dokumentation von dieser Entwicklung nicht profitieren.

Im Jahr 2009 untersuchten Fiederer & O'Brien den KS-Einfluss im Zusammenhang mit einem RBMÜ-System nach den Definitionen von Hutchins und Somers ebenfalls unter Berücksichtigung der Idiomatik der MÜ in der technischen Dokumentation (vgl. Fiederer & O'Brien 2009). Mit der vorliegenden Studie, wird die KS und ihr Einfluss auf die maschinelle Übersetzbarkeit in Anbetracht der jüngsten MÜ-Entwicklung nach denselben Qualitätsdefinitionen in einem Vergleich von RBMÜ-, SMÜ-, HMÜ- und NMÜ-Systemen empirisch wieder aufgegriffen, um die technologische Entwicklung der MÜ zu verfolgen und Implikationen für die technische Dokumentation zu erarbeiten.

(2) *Die Eignung der MÜ für die Intention des Inhalts*: Bei dem zweiten Qualitätskriterium geht es um die Eignung der MÜ für die Intention des Inhalts. Nach der Sprechakttheorie kann eine Aussage mit mehr als einer Illokution verbunden sein (vgl. Rehbein 1988). In der technischen Dokumentation gilt gleichzeitig die Eindeutigkeit der intendierten Aussage als „wesentliches Güterkriterium" (Lehrndorfer 1996b: 156). Daher ist „die Reflexion der intendierten Pragmatik der Aussage und ihre Benennung durch den technischen Redakteur von Bedeutung" (ebd.). Nach Lehrndorfer (ebd.: 155) soll „der technische Redakteur zunächst eine genaue Zielsetzung seiner Aussage reflektieren" und die jeweilige Aussageintention vor jeder Satzeinheit als Auszeichnungsformat nennen bzw. annotieren.

In der technischen Dokumentation existieren mehrere Klassifizierungen für die Aussageintentionen. Lehrndorfer (ebd.: 155ff.) unterscheidet zwischen den drei Hauptkategorien Handlungsanweisung, Sicherheitshinweis und Produktbeschreibung, denen weitere Unterkategorien zugeordnet sind. Eine weitere Klassifizierung der Aussageintentionen in der technischen Dokumentation bietet Ley (2005: 110ff.). Diese umfasst die Illokutionen Assertion, Direktiv und Expressiv, denen ebenfalls detaillierte Unterkategorien zugeordnet sind (ebd.). Auf dieser Basis war es essentiell, die Eignung der Übersetzung für die Intention des Inhalts zu evaluieren.

(3) *Die korrekte bzw. klare orthografische Darstellung der MÜ*: Hierbei wird die Orthografie als grafisches Mittel für eine adäquate Darstellung des Inhalts miteinbezogen, was wiederum der Verständlichkeit, Genauigkeit sowie der Vermittlung seiner Intention dienlich ist. Nach Nerius (2007: 30f.) ist die Orthografie (Rechtschreibung)[51] die „Norm der formalen Seite der geschriebenen Sprache, und zwar aller Teilbereiche der Schreibung einschließlich der Interpunktion". Durch die grafischen Mittel der Orthografie werden unterschiedliche Eigenschaften der Sprachelemente verdeutlicht (ebd.: 91). Vier grafische Mittel der Orthografie waren für die Studie von besonderer Bedeutung:

- Die Anführungszeichen: „Anführungszeichen haben die besondere Fähigkeit, bestimmte kommunikative Bezüge herzustellen und damit eine besondere Aussageabsicht des Schreibenden zum Ausdruck zu bringen." Dementsprechend – anders als das paarige Komma – ist deren Aufgabe nicht bloß „die Kennzeichnung der Grenzen eines Einschubs innerhalb des Satzverbandes", sondern vielmehr „die besondere Charakterisierung dieses Einschubs", so markieren die Anführungszeichen fremde Äußerungen wie Zitate, Titel und Überschriften (Nerius 2007: 253f.), vgl. insb. Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" unter §5.4.2.
- Der Ergänzungsstrich ist eine Art des Bindestriches, der eine ökonomische (ebd.: 187) und zugleich eine semantische Funktion hat (ebd.: 191). Mithilfe des Ergänzungsstriches wird eine Auslassung markiert (ebd.: 187) und die „Zusammengehörigkeit räumlich getrennter Bestandteile zusammengesetzter oder abgeleiteter koordinierter Wörter innerhalb der Wortgruppe" signalisiert (ebd.: 191). Somit dient der

[51] Orthografie kommt aus dem Griechischen; „orthós" für „recht" und „gráphein" für „schreiben" (Nerius 2007: 30).

Ergänzungsstrich der „Monosemierung“ (d. h. der Eindeutigkeit) „der Aussage im Interesse des Lesenden und wirkt damit der Gefahr eines Missverständnisses entgegen“ (vgl. fehlender Ergänzungsstrich im Beispiel „Max Müller, Pinsel und Bürstenmachermeister“). Es gibt keine Regel zur Verwendung bzw. Nicht-Verwendung des Ergänzungsstriches. Die Entscheidung über seine Verwendung liegt bei dem Schreibenden. Entscheidet er keine Wortteile wegzulassen, kann dies „zwar gegen stilistische Normen verstoßen, stellt aber keinen orthographischen Fehler dar“. (ebd.) Vgl. insb. Regel „Keine Wortteile weglassen“ unter §5.4.10.

- Die Großschreibung, wodurch die Wortklasse der Substative sowie der Anfang eines Satzes bzw. einer Bezeichnung gekennzeichnet wird (ebd.: 91), vgl. insb. Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ unter §5.4.2.
- Das paarige Komma, das „eine Klammerfunktion hat, indem es syntaktische Einheiten einschließt und dadurch aus dem übrigen Satzverband heraushebt“ (ebd.: 252), vgl. insb. Regel „Partizipialkonstruktionen vermeiden“ unter §5.4.6.

In seinem Werk „Deutsche Orthographie“ beleuchtet Nerius (2007: 91), wie durch die grafischen Mittel der Orthografie Aspekte der Satzbedeutung auf der grafischen Ebene gekennzeichnet werden und somit das Erkennen der Satzform erleichtert und die rasche Erfassung der Satzbedeutung unterstützt wird. Zudem zeigt Nerius (ebd.: 275ff.), die Rolle, die die grafischen Mittel der Orthografie bei der stilistischen Differenzierung spielen können.[52]

Vor diesem Hintergrund ist zu erwarten, dass eine klare bzw. korrekte orthografische Darstellung des Inhalts einen Einfluss auf die MÜ-Qualität hat; durch die stilistische Differenzierung, die sie bewirkt, beeinflusst sie die Verständlichkeit und die Genauigkeit der Aussage. Dementsprechend wurde die klare orthografische Darstellung als Kriterium der Stilqualität betrachtet, letztendlich ist die Orthografie, wie ihre Definition besagt (Nerius 2007: 30f.), die Norm der „formalen“ Seite der geschriebenen Sprache.

Anhand des Stilqualität-Scores auf der Likert-Skala wurde die Stilqualität operationalisiert und entsprechend wurde der Einfluss des Einsatzes der KS-Regeln

[52] Der Effekt der Orthografie z. B. bei der Kommasetzung wird von Nerius (2007: 238) anhand von Beispielen wie „Peter versprach, der Mutter zu schreiben“ vs. „Peter versprach der Mutter, zu schreiben“ und „Der Kranke drohte, sich ein Leid anzutun“ vs. „Der Kranke drohte sich ein Leid anzutun“ verdeutlicht.

auf die Stilqualität quantitativ bewertet. Analog zur Bewertung der Inhaltsqualität wurden die Bewerter in den Testanweisungen aufgefordert, erst die relevanten Qualitätskriterien (siehe [3a] in Abbildung 4.6) anzukreuzen und ihre Kriterienauswahl zu begründen ([4] in Abbildung 4.6) und danach den Score der Stilqualität auf Basis der ausgewählten Kriterien zu vergeben (siehe „Testaufgaben“ unter §4.5.5.2); Ziel dieser Vorgehensweise war, sicherzustellen, dass die Bewertungen der drei genannten Kriterien in den Score der Stilqualität einfließen.

4.5.5.2 Testlayout

In diesem Abschnitt wird das Testlayout genau beschrieben; dies umfasst die Klärung, wie die Evaluationsfragen gestellt wurden, welche Bewertungsskala verwendet wurde, wie die Ausgangssätze, die Zielsätze sowie die Evaluation im Allgemeinen den Teilnehmern präsentiert wurden und wie der Evaluationsablauf sich gestaltete.

4.5.5.2.1 Form der Evaluationsfragen

Wie im Kapitel MÜ (unter §3.4.3.1) detailliert dargestellt, sind Ranking und Scoring die dominierenden Bewertungsmethoden im Bereich der Humanevaluation. Unter Berücksichtigung der Vor- und Nachteile der beiden Methoden, wurde in der vorliegenden Studie *Scoring* angewendet. Im Folgenden werden die Gründe dieser Wahl näher erläutert.

In der Humanevaluation der Studie war ein Ranking der Übersetzungen aufgrund der großen Anzahl der MÜ-Sätze pro Ausgangssatz nicht möglich. Auch trotz der Einschränkung der Anzahl der MÜ-Sätze auf 5–6 Übersetzungen pro Ausgangssatz, bleibt die Auswertung der Rankingergebnisse von fehlerhaften MÜ-Sätzen problematisch. Betrachten wir ein Beispiel, in dem zwei Ausgangssätze jeweils fünfmal übersetzt wurden: Die Übersetzungen vom ersten Ausgangssatz umfassen zwei fehlerfreie Übersetzungen und drei fehlerhafte Übersetzungen mit *wenig gewichtigen Fehlern*. Bei dem zweiten Ausgangssatz gibt es ebenfalls zwei fehlerfreie Übersetzungen und drei fehlerhafte Übersetzungen, allerdings mit *gewichtigen Fehlern*. Bei den beiden Ausgangssätzen erzielen die fehlerhaften Übersetzungen vergleichbare schlechte Rankings trotz des unterschiedlichen Schweregrads der Fehler (vgl. Costa u. a. 2015).

Der einzige Fall, in dem ein Ranking trotz der Schwierigkeit des Rankings mehrerer Sätze vorteilhaft sein kann, ist das Ranking der MÜ-Sätze der Annotationsgruppe (RR). In dieser Annotationsgruppe sind die MÜ-Sätze sowohl vor

als auch nach dem Einsatz der KS-Regeln fehlerfrei. Beim Scoring von solchen Fällen könnten die Bewerter dazu neigen, die Sätze schnell – ohne Beachtung von stilistischen Feinheiten – mit 5 Punkten (beste Bewertung) zu bewerten. Durch das Ranking könnte eine solche Bewertung vermieden werden, da die Bewerter die MÜ-Sätze genauer betrachten und vergleichen müssen. Jedoch konnte die Gruppe RR nicht gesondert gerankt werden, denn eine Mischung der beiden Bewertungsmethoden (Scoring bei fehlerhaften Übersetzungen und Ranking bei fehlerfreien Übersetzungen) wäre bei der Auswertung problematisch. Für manche Sätze wiederholten sich die MÜ z. B. bei zwei Systemen, wobei die MÜ bei einem System innerhalb der Annotationsgruppe RR und bei dem anderen innerhalb der Annotationsgruppe FR (falsche MÜ vor KS; richtige MÜ nach KS) war. In solchen Fällen und bei der Verwendung von zwei Bewertungsmethoden hätte eine und dieselbe Übersetzung zwei Bewertungen jeweils nach einer Methode.

Außerdem bestehen weitere Schwierigkeiten beim Ranking: Erstens, ein Ranking kann auch dazu führen, dass die Teilnehmer mehrere Übersetzungen eines Satzes identisch bewerten, es sei denn, die Testanweisungen verbieten identische Bewertungen, sog. „ties“. Dies bildet eine weitere Schwierigkeit für die Teilnehmer, da oft unterschiedliche Fehlertypen schwer vergleichbar sind, sodass sie in Relation zueinander gerankt werden können. Beispielsweise lassen sich bei der Regel „Anführungszeichen verwenden“ die Übersetzungsvarianten Kleinschreibung sowie Großschreibung mit und ohne Anführungszeichen schwer vergleichen und entsprechend ranken. Aus diesen Gründen wurden die MÜ-Sätze mithilfe einer 5-Punkte-Likert-Skala bewertet.

4.5.5.2.2 Bewertungsskala

Skalen, wie die Likert-Skala, sind ein weit verbreitetes Bewertungsinstrument im Bereich der MÜ-Evaluation. Die Spanne der Likert-Skala liegt in der Regel zwischen fünf und neun Punkten (vgl. Porst 2011; Saldanha & O’Brien 2014: 157). Einige Forscher bevorzugen der Verwendung von Skalen mit geraden Zahlen, um zu verhindern, dass die Teilnehmer vermehrt den Mittelwert der Skala wählen. Auf diese Weise versuchen sie dem Risiko entgegenzuwirken, dass das Endergebnis unentschieden bleibt, was wiederum eine Nichtbeantwortung der Forschungsfrage zur Konsequenz haben könnte. (Saldanha & O’Brien 2014: 157f.; Johnston 2015) Gleichzeitig hat die Verwendung von Skalen mit geraden Zahlen den Nachteil, dass die Teilnehmer keine Möglichkeit haben, die tatsächlich neutralen Fälle entsprechend zu bewerten. Das kann wiederum die Ergebnisse verzerren und die Teilnehmer frustrieren. (Johnston 2015) Die unter §3.4.1 „Qualität der MÜ“ dargestellten Studien führten die Bewertung mithilfe einer Likert-Skala mit vier (Van

Slype 1979; Vanni & Miller 2002; Coughlin 2003) oder fünf Punkten (LDC 2002; White 2003; Hamon 2007) durch.

In der vorliegenden Studie wurden die MÜ-Sätze auf einer 5-Punkte-Likert-Skala bewertet. Die Spanne der Skala war für den erzielten Differenzierungsgrad angemessen. Eine kürzere Skala wäre für diese Studie ungeeignet, da die Erfassung von kleinen Qualitätsdifferenzen für den Vergleich der Szenarien vor-KS vs. nach-KS notwendig ist. Eine noch längere bzw. ausdifferenziertere Skala würde die Ratingkonsistenz beeinträchtigen. Ferner unterstützte die ortsflexible Durchführung der Evaluation über einen Zeitraum von drei bis vier Wochen die Teilnehmer dabei, flexibel und ohne Zeitdruck die Übersetzungen ausdifferenziert zu bewerten, was wiederum dabei unterstützend wirkte, das Risiko einer vermehrten Wahl des Mittelwerts der Skala zu minimieren.

4.5.5.2.3 Darstellung der Ausgangssätze

Bei der Erstellung des Testdesigns stellte sich die Frage, ob die Darstellung des Ausgangstexts erforderlich ist. Eine Darstellung des Ausgangssatzes ist notwendig, damit die Genauigkeit der Übersetzung beurteilt werden kann (z. B. kann der lexikalische Fehler „unübersetztes Wort" nur durch eine Darstellung des Ausgangssatzes erkannt werden). Auf der anderen Seite können die Verständlichkeit und der Stil ohne Ausgangstext bewertet werden. Diese Hypothese bestätigte White (2003: 205), als er zunächst die Qualität von MÜ-Sätzen im Allgemeinen (im Sinne von „good English", „degraded by up ton errors" or „wrong") bewertete. Er kam zum Ergebnis, dass ohne eine Darstellung des Ausgangstexts „we do not know anything about where these expressions came from. Are they really translations of anything?" (ebd.). White (ebd.) ging diese Problematik an, indem er den Ausgangstext den Bewertern zur Verfügung stellte. Dies ermöglichte ihm, die Qualität detailliert gemäß den Attributen „Intelligibility" und „Fidelity" zu untersuchen und in der Lage zu sein, „to tell something about the translation issues from looking at both the source and target language" (ebd.: 206). Er nannte gleichzeitig die damit verbundene Schwierigkeit, den Test von Übersetzern durchführen lassen zu müssen, die möglicherweise schwer zu einer Teilnahme zu motivieren sind (ebd.).

Im Vergleich zu Whites Studie sind in der vorliegenden Studie nicht nur die „translation issues" von Interesse, sondern auch die fehlerfreien Fälle. Eine feine Differenzierung im Stil, Verständlichkeits- und Genauigkeitsgrad von fehlerfreien Übersetzungen muss auch abgedeckt werden und ist – vor allem aufgrund der großen Anzahl von fehlerfreien Übersetzungen – von großer Bedeutung, damit ein umfangreicher Vergleich der vor- und nach-KS-Szenarien gewährleistet

werden kann.

Wie oben erwähnt, ist eine Darstellung des Ausgangssatzes zur Beurteilung der Genauigkeit (und somit der Inhaltsqualität) erforderlich. Für die Stilqualität wurde die Entscheidung, ob die Ausgangssätze den Bewertern zur Verfügung gestellt werden oder nicht, nach der Durchführung eines Testlaufs getroffen. Der Testlauf bestand aus zwei Phasen: In der ersten Phase erhielten drei Teilnehmer 14 Sätze ohne den Ausgangssatz und hatten die Aufgabe die Stilqualität auf einer 5-Punkte-Likert-Skala zu bewerten und etwaige Fehler zu kommentieren (Abbildung 4.4).[53]

	very low	low	acceptable	high	very high
Only connect the device to a correctly installed plug socket.	[ ]	[ ]	[ ]	[ ]	[ ]
Comment:					

Style quality: The extent to which the translation flows well and sounds natural and idiomatic in Standard Written English.

Abbildung 4.4: Beispiel aus der ersten Phase des Testlaufs

Ein Monat später fand die zweite Phase statt. Hier haben dieselben drei Teilnehmer die Stil- und Inhaltsqualität derselben 14 Sätze bewertet, wobei die Ausgangssätze zur Verfügung standen (Abbildung 4.5).

Das Gerät nur an eine vorschriftsmäßig installierte Steckdose anschließen.
Only connect the device to a correctly installed plug socket.

	very low	low	acceptable	high	very high
Content quality	[ ]	[ ]	[ ]	[ ]	[ ]
Style quality	[ ]	[ ]	[ ]	[ ]	[ ]

Comment:

Content quality: The extent to which the translation reflects the information in the source text accurately and in an easy way to understand.
Style quality: The extent to which the translation flows well and sounds natural and idiomatic in Standard Written English.

Abbildung 4.5: Beispiel aus der zweiten Phase des Testlaufs

Ein Vergleich der Ergebnisse der Stilqualität in den beiden Phasen zeigt, dass die Bewertung von 57 % der Sätze unverändert blieb, von 28 % der Sätze sich

[53] Im Testlauf wurden die Qualitätsdefinitionen vereinfachter bzw. kürzer (als unter §4.5.5.1) angegeben, da es im Test primär um die Darstellung der Ausgangssätze ging.

um einen Punkt auf der Skala veränderte und von 15 % der Sätze sich um mehr als einen Punkt auf der Skala veränderte. Zudem zeigten die Kommentare der Teilnehmer, dass der Ausgangstext bei der Bewertung des Stils eine Rolle spielte. Beispielsweise empfahlen in der zweiten Phase (Ausgangssatz dargestellt) zwei Bewerter einen Passivsatz ins Aktiv zu übersetzen. Auch auf semantischer Ebene war es hilfreich, den Ausgangssatz einzublenden, z. B. wurde in einem Satz ‚Mängel berichten' von einigen Systemen als ‚announce defects' übersetzt. Zwei Teilnehmer schlugen anhand des Ausgangssatzes ‚report defects' zur Verbesserung der Stil- und Inhaltsqualität vor.

Auf Basis dieser Ergebnisse und aufgrund der großen Anzahl der zu bewertenden MÜ-Sätze (1.100 Sätze) wurde entschieden, die Ausgangsätze bei der Bewertung einzublenden. Eine Bewertung der Inhaltsqualität ohne den Ausgangstext wäre nicht möglich. Eine Aufteilung der Sätze in zwei Phasen, wobei die Teilnehmer in der ersten Phase die Stilqualität ohne Ausgangstext bewerten und nach einem bestimmten Zeitraum die Inhaltsqualität mit Ausgangstext in einer zweiten Phase bewerten, hätte das Finden von Teilnehmern erschwert, die für einen längeren Zeitraum verfügbar gewesen wären, und wäre zudem mit weiteren Kosten verbunden gewesen.

4.5.5.2.4 Darstellung der Zielsätze: Markierung der KS-Stelle

Bei dem Testdesign stellte sich die Frage, ob es für die Bewertung sinnvoll wäre, die KS-Stelle zu markieren und die Teilnehmer aufzufordern, sich bei der Bewertung gezielt auf die markierte Stelle zu beziehen und damit zu verhindern, dass sie von anderen Stellen außerhalb der KS-Stelle abgelenkt werden. Die Technik der Markierung einer bestimmten Stelle innerhalb der MÜ wurde von Callison-Burch u. a. (2007) und Ramírez Polo (2012: 236) bei der Humanevaluation eingesetzt. Callison-Burch u. a. (2007) markierten einen bestimmten syntaktischen Bestandteil (im Ausgangs- und Zielsatz) und forderten die Teilnehmer auf, mehrere MÜ eines Ausgangssatzes zu ranken. Bei diesem Ranking berichteten Callison-Burch et al. (ebd.) von einem hohem Interrater- und Intrarater-Agreement sowie schneller Evaluation. In der vorliegenden Studie handelt es sich hingegen um eine Scoring-Aufgabe von einer MÜ und keiner Ranking-Aufgabe von mehreren MÜ desselben Satzes. Ramírez Polo (2012: 236) markierte die KS-Stelle (im Ausgangs- und Zielsatz) und bat die Teilnehmer die MÜ auf einer 4er-Skala zu bewerten. Sie beschrieb das Profil der Teilnehmer als „native speakers who worked within the automotive industry and therefore mastered the terminology and the background knowledge necessary to understand the texts" (ebd.). Aus dieser Profilbeschreibung geht hervor, dass die Teilnehmer erfahrene Beschäftigte der

Automobilindustrie sind. Informationen zu ihren linguistischen oder translatorischen Qualifikationen kann man jedoch nicht erschließen. Ramírez Polo macht ebenfalls Gebrauch von der Markierung der KS-Stelle im Ausgangs- und Zielsatz, „in order to direct the evaluator's attention to these fragments" (ebd.). Die Markierung diente somit der Orientierung, die sich eventuell für ihre Zielgruppe als nützlich erwies. Anders als Ramírez Polos Studie wurde die vorliegende Studie von qualifizierten Übersetzern durchgeführt (siehe §4.5.5.3).

Um eine Entscheidung diesbezüglich zu treffen, wurden drei Tests mit insgesamt zehn Teilnehmern durchgeführt: Im ersten Test *Evaluation ohne Markierung* wurden 19 Sätze von vier Teilnehmern ohne Markierung der KS-Stelle bewertet. Im zweiten Test *Evaluation mit langer Markierung* wurden dieselben 19 Sätze von drei Teilnehmern bewertet, nachdem die KS-Stelle zusammen mit anderen Wörtern im Satz markiert wurde. Die Länge der Markierung betrug min. 50 % des Satzes gemessen auf Basis der Anzahl der Wörter. Im dritten Test *Evaluation mit kurzer Markierung* wurden ebenfalls die KS-Stellen in den 19 Sätzen markiert, aber mit einer kürzeren Markierung, so dass die KS-Stelle zusammen mit zwei weiteren Wörtern markiert wurde (d. h. mit einem Wort vor und einem Wort nach der KS-Stelle, wenn die KS-Stelle in der Mitte des Satzes lag; oder mit zwei Wörtern nach der KS-Stelle, wenn die KS-Stelle am Anfang des Satzes lag). Die Sätze wurden ebenfalls im letzten Test von drei Teilnehmern bewertet. Die Ergebnisse der drei Tests lauten wie folgt:

Tabelle 4.12: Ergebnisse des Testlaufs

	KS-Stelle ohne Markierung	KS-Stelle mit langer Markierung	KS-Stelle mit kurzer Markierung
Mittelwert der Stilqualität	3,80	4,00	3,96
Mittelwert der Inhaltsqualität	4,44	4,33	4,49

Wie Tabelle 4.12 zeigt, liegt die Differenz im Mittelwert bei dem Szenario „ohne Markierung" gegenüber dem Szenario „kurze Markierung" bei 3,9 % für die Stilqualität bzw. bei 1 % für die Inhaltsqualität. Zudem kommentierten die Teilnehmer die KS-Stelle wie folgt: Im ersten Test (*Evaluation ohne Markierung*) konnten die Teilnehmer die KS-Stelle in 74 % der Sätze identifizieren und kommentierten

sie. In ca. 16 % der Fälle kommentierten die Teilnehmer die KS-Stelle zusammen mit einer weiteren Stelle. In den restlichen Fällen (10 %) waren die MÜ fehlerfrei, sodass kein Kommentar erforderlich war. Auf der anderen Seite zeigten die Tests mit langer und kurzer Markierung, dass die Teilnehmer Schwierigkeiten hatten, sich bei der Qualitätsbewertung auf einen bestimmten Teil der Übersetzung zu beschränken und einen Qualität-Score in Bezug auf die markierte Stelle zu vergeben. Trotz der Testanweisung, nach der nur die markierte Stelle bei der Bewertung berücksichtigt werden soll, beinhalteten die Kommentare häufig Stellen, die nicht markiert waren. Dies kann auch je nach Satzstruktur, den enthaltenen Konstruktionen und der Art der Fehler nachvollziehbar sein.

Aufgrund des kleinen Unterschieds in den Mittelwerten (Tabelle 4.12) in den Szenarien mit und ohne Markierung sowie der beschriebenen Bewertungsschwierigkeit wurde entschieden, die Übersetzungen bei der Evaluation ohne Markierung darzustellen. Im Gegensatz zu den oben genannten Studien, die die Markierung zur Orientierung einsetzen, sieht die vorliegende Studie eine Gefahr in der Markierung, nämlich dass die Teilnehmer durch die Markierung einen Anreiz erhalten, eine Stelle zu kritisieren, die sie ohne Markierung nicht bemängelt hätten. Diese Gefahr wurde durch das Testing ohne Markierung abgewendet.

4.5.5.2.5 Darstellung der Evaluation

Nachdem die Qualitätsaufteilung und -kriterien sowie die Bewertungsskala festgelegt wurden, folgte eine Präzisierung der Details des Testdesigns. Diese beziehen sich auf die genauen Testanweisungen und -aufgaben sowie das Testlayout.

Testlayout Wie Abbildung 4.6 zeigt, bestand das Testlayout aus fünf Bereichen: Im Bereich [1] stehen die Ausgangs- und Zielsätze; im Bereich [2] die Likert-Skalen für die Stil- und Inhaltsqualität; in den Bereichen [3a] und [3b] die Kriterien der Stil- bzw. Inhaltsqualität; im Bereich [4] Kommentarfelder; und im Bereich [5] ein Feld für den Vorschlag einer Alternativübersetzung.

Testaufgaben Die Teilnehmer hatten *drei Aufgaben*, die nach den Testanweisungen (Abbildung 4.7) in der folgenden *Reihenfolge* beantwortet werden sollten:

(1) *Auswahl der zutreffenden Kriterien der Stil- und Inhaltsqualität* (Bereich [3a] und [3b] in Abbildung 4.6). Da mehrere Kriterien gleichzeitig zutreffen bzw. sich gegenseitig beeinflussen können, wurden für die einzelnen Kriterien Checkboxen verwendet, die die Auswahl mehrerer Kriterien erlauben.

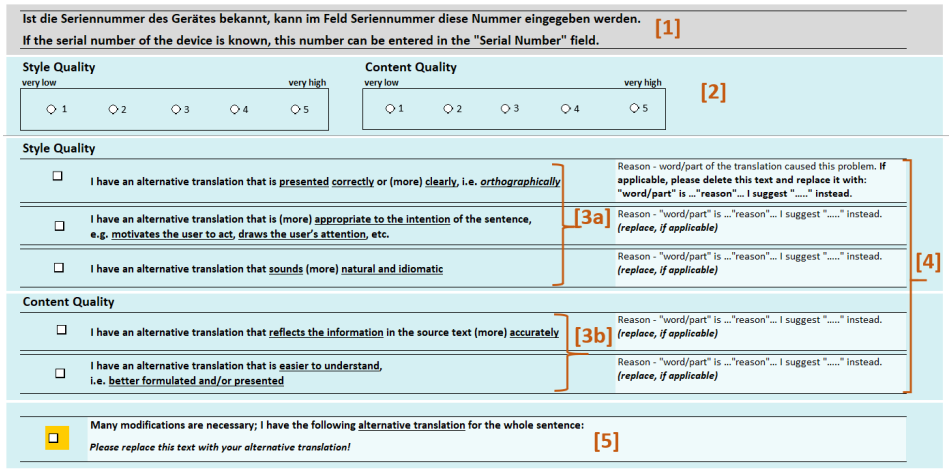


Abbildung 4.6: Testlayout der Humanevaluation

(2) *Begründung der Auswahl bestimmter Qualitätskriterien* (Bereich [4] in Abbildung 4.6), indem der Bewerter den für das ausgewählte Kriterium relevanten Teil der Übersetzung angibt, seine Auswahl kurz begründet und eine Korrektur bzw. mögliche Verbesserung vorschlägt. Sollten mehrere Modifikationen erforderlich gewesen sein, gibt der Bewerter eine Alternativübersetzung für den gesamten Satz in Bereich [5] an.

(3) *Vergabe der Stil- und Inhaltsqualität-Scores* entsprechend der ausgewählten Qualitätskriterien auf den Likert-Skalen (Bereich [2] in Abbildung 4.6).

Testanweisungen Die Bewerter erhielten zu Beginn der Testphase eine kurze Einführung zur Evaluation, die die allgemeinen Evaluationsanweisungen, die Definitionen der Qualität sowie eine Erklärung des Testlayouts anhand eines Beispiels (siehe Anhang A) beinhaltete. Es wird in der Regel empfohlen, die Definitionen der Qualität am Ende jeder Seite anzugeben, die als Gedächtnisstütze dienen sollen (Doherty 2012: 193). Die alleinige Angabe der Definitionen birgt das Risiko, dass sie nicht gelesen bzw. umgesetzt werden. Anstatt die Definitionen bei jedem Satz isoliert anzugeben, wurden sie in die anzukreuzenden Kriterien (Bereich [3a] und [3b] in Abbildung 4.6) integriert. Dies zusammen mit der darauffolgenden Aufgabe, die angekreuzten Kriterien zu kommentieren (Bereich [4] in Abbildung 4.6), zielte darauf ab, dass die Evaluationsdefinitionen umgesetzt werden.

Der Test wurde in Microsoft Excel erstellt,[54] so dass die erste Seite die Anweisungen beinhaltet, gefolgt von 25 Seiten zur Bewertung jeweils eines Satzes (Abbildung 4.7).

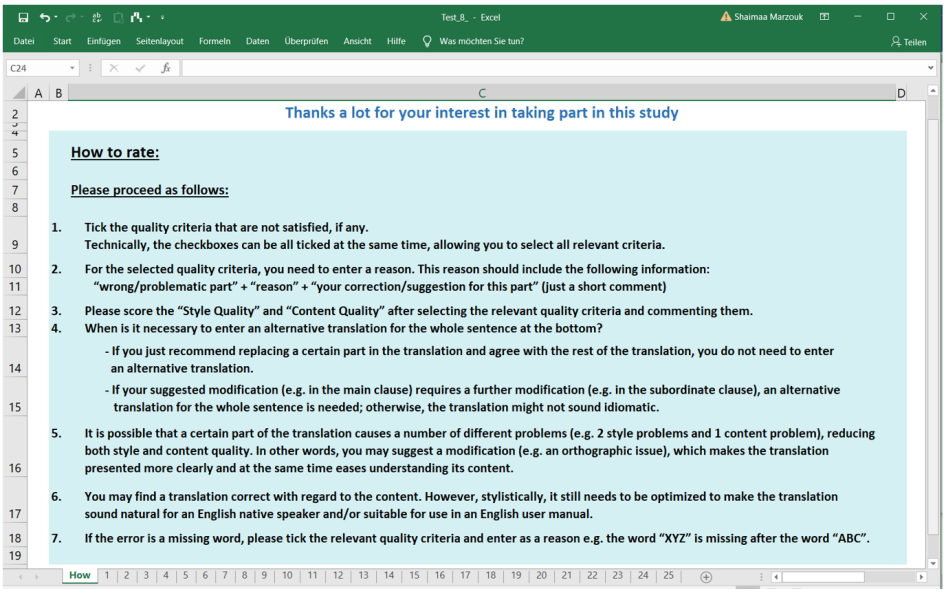


Abbildung 4.7: Testanweisungen und Testlayout in Microsoft Excel

In den Testanweisungen wurden die Bewerter darauf aufmerksam gemacht, dass (1) ein bestimmter Teil der Übersetzung einen Einfluss auf mehrere Qualitätskriterien haben kann (1:n-Beziehung möglich), siehe Punkt 5 in den Anweisungen; und (2) eine inhaltlich korrekte Übersetzung stilistisch optimierungsbedürftig sein kann (Punkt 6 in den Anweisungen).

Die Testanweisungen beeinflussen zweifellos die Studienvalidität, -reliabilität und Evaluationseffizienz (Doherty 2017). Nach Doherty (ebd.: 11) „Instructions and guidelines should be written clearly and concisely and contain explicitly operationalized definitions appropriate to the evaluator group and the TQA[55] task". All diese Faktoren wurden bei dem Testdesign berücksichtigt. Die Klarheit der Testanweisungen bzw. -aufgaben wurden in einem kurzen Testlauf mit drei Probanden[56] geprüft. Auf Basis dieses Testablaufs wurden die notwendigen An-

[54] Aufgrund des speziellen Aufbaus der Testaufgaben war eine Durchführung des Tests mit einem klassischen Testtool nicht möglich. Mithilfe von Microsoft Excel konnte das gewünschte Design entworfen werden und die Errechnung sämtlicher quantitativer Daten war realisierbar.

[55] TQA steht für Translation Quality Assessment.

[56] Die Probanden des Testlaufs nahmen nicht an der Evaluation teil.

passungen vorgenommen. Die Anzahl der Sätze pro Test wurde auf Basis des Testlaufs festgelegt. Im Testlauf nahm die Bewertung eines Satzes ca. 2 Minuten in Anspruch. Um die Testdauer in einem Zeitraum von weniger als einer Stunde zu halten und somit die Auswirkungen von Müdigkeit oder Langeweile zu reduzieren, wurde die Anzahl der Sätze pro Test auf 25 festgelegt. Auf Seite 1 von 25 und Seite 25 von 25 musste jeder Bewerter die Startzeit bzw. Endzeit angeben. Nach den Bewertungsseiten folgte eine für die Bewerter ausgeblendete Seite, in der alle Ergebnisse sowie Zeitangaben mithilfe von Excel-Formeln zusammengefasst wurden.

Ablauf der Evaluation Die 1.100 Testsätze (siehe Anhang C) wurden randomisiert in 44 Tests aufgeteilt. Jeder Bewerter hatte die Möglichkeit, je nach seiner Kapazität festzulegen, ob er einen, zwei oder drei Tests pro Tag bewerten wollte. Grundvoraussetzung war, mindestens einen Test täglich zu bewerten, und somit Unterbrechungen zu vermeiden, die möglicherweise das Intrarater-Agreement negativ beeinflusst hätten. Das Maximum wurde auf drei Tests täglich festgelegt, um möglichst eine Bewertung trotz Ermüdung zu verhindern. Außerdem wurden die Teilnehmer gebeten, eine Pause zwischen den Tests einzulegen. Die Bewerter erhielten die Tests in unterschiedlich randomisierter Reihenfolge (z. B. erhielt Bewerter 1 Test 15, Test 8, Test 40 hintereinander). Der Tester erhielt jeden Tag die bewerteten Tests und prüfte, ob alle Sätze bewertet und ggf. kommentiert worden waren. Bei fehlenden Bewertungen wurde der Teilnehmer gebeten, sie zu ergänzen. Daraufhin erhielt der Teilnehmer die Tests des folgenden Tages. Dieser Ablauf hatte zum Ziel, die Qualität der Bewertung zu sichern. Als Vergütung erhielten die Teilnehmer 25 Cent pro Satz.

Psychologische Risiken, die mit einer Humanevaluation einhergehen, lassen sich nur schwer vollständig verhindern. Durch die zweifache Randomisierung konnte der Effekt einiger Risiken minimiert werden: Erstens ähneln sich die MÜ-Sätze verschiedener Systeme desselben Ausgangssatzes zu einem gewissen Grad. Eine Randomisierung der Testsätze zielte darauf ab, Qualitätsbewertungen zwischen benachbarten Sätzen so unabhängig wie möglich zu halten (vgl. Hamon 2007). Zweitens lief die Testphase über drei bis vier Wochen. Eine Randomisierung der Reihenfolge der von jedem Teilnehmer beantworteten Tests hatte zum Ziel, das Risiko, das von White (2003) als „Maturation“ bezeichnet wird, zu mitigieren. Im Laufe der Testphase „ordinary things can affect someone's ability to be consistent in their judgments. Specifically, they will get tired, bored, hungry, or fed up with the process of evaluating” (ebd.: 209). Das hat den Effekt, dass „sentences they graded later in the cycle will get a different look than the ones they

graded earlier“ (ebd.). Dadurch, dass die Bewerter die Tests in unterschiedlicher Reihenfolge erhielten, wurden keine identischen Sätze von allen Teilnehmern am Ende der Testphase bewertet bzw. wurden die Sätze von den Teilnehmern zu unterschiedlichen Zeitpunkten in Laufe der Testphase bewertet.

4.5.5.3 Anzahl und Profil der Teilnehmer

Vorherige Studien der MÜ-Evaluation empfehlen die Evaluation mithilfe von mindestens drei bzw. vier Probanden durchzuführen (vgl. Dyson & Hannah 1987: 166; Arnold 1994: 171). Nach dem Gesetz des abnehmenden Grenznutzens (law of diminishing marginal utility) von Gossen (1854/1983 zit. nach Caplin & Glimcher 2014) „each additional unit of gain leads to an ever-smaller increase in subjective value“. Eine Anwendung dieses bewährten Konzepts im Experimentkontext würde Folgendes bedeuten: Mit der Erhöhung der Anzahl der Teilnehmer liefert jeder neue Teilnehmer weniger neuen Input als der vorherige Teilnehmer (vgl. Huber 2008: 37). Dies wiederum deutet darauf hin, dass „at some point, additional sampling serves little purpose since these measurements provide negligible new information. The optimal sample size falls in this region of diminished marginal utility” (ebd.).

In dieser Studie wurde das Gesetz des abnehmenden Grenznutzens zur Festlegung der Teilnehmeranzahl herangezogen. In der Testphase wurden zunächst fünf Teilnehmer rekrutiert. Sukzessive wurde die Anzahl der Teilnehmer erhöht bis sich der Mittelwert der akkumulativen Qualitätswerte stabilisierte. Bei acht Teilnehmern veränderte sich der akkumulative Qualitätsmittelwert kaum. Entsprechend wurde die Anzahl der Teilnehmer nicht weiter erhöht.

Die Hauptauswahlkriterien für die Teilnehmer waren:

(1) Englischmuttersprachler: Dies stellte das Hauptkriterium dar, da die Zielsprache der MÜ Englisch war.

(2) Bachelorstudiengang der Translation bereits abgeschlossen. Mit diesem Kriterium wurden hohe Sprachkompetenz in der deutschen Sprache sowie Translationskompetenz sichergestellt.

(3) Masterstudent im letzten Semester des Masterstudiengangs Translation des Fachbereiches Translations-, Sprach- und Kulturwissenschaft der Johannes Gutenberg-Universität Mainz.

Im Rahmen der Evaluationsdesignphase wurden Professionals und Semiprofessionals kontaktiert. Im Vergleich zu Professionals waren die Semiprofessio-

nals einfacher zu erreichen, zeigten eine höhere Bereitschaft über mehrere Wochen verfügbar zu sein und ihre Vergütung war kosteneffizienter. Daher wurde die Studie mithilfe von Semiprofessionals durchgeführt. Dieses Profil hat Vorteile gegenüber einer Evaluation mithilfe von Endnutzern oder Professionals: Endnutzer verfügen zwangsläufig nicht über die linguistischen Kompetenzen, die für die Evaluation erforderlich sind (vgl. Roturier 2006: 201). Vorherige MÜ-Evaluationsstudien (vgl. Arnold 1994; Fiederer & O'Brien 2009), die den Unterschied zwischen der Qualitätsbewertung von Übersetzern gegenüber der von MÜ-System-Endnutzern adressieren, fanden heraus, dass die Bewertung der Übersetzer aufgrund ihrer linguistischen bzw. translatorischen Qualifikationen strenger und kritischer im Vergleich zu der der Endnutzer ist. Für die Humanevaluation der vorliegenden Studie ist eine strenge bzw. kritische Bewertung der MÜ-Qualität von Vorteil, da die Aufgabenstellung eine Differenzierung zusammen mit einer genauen Begründung der identifizierten Fehler verlangt. Gleichzeitig sind die Teilnehmer dieser Studie Übersetzer mit relativ kurzer Übersetzungserfahrung, daher ist eine angemessen kritische Bewertung zu erwarten.

Die Profildaten der Teilnehmer wurden im Rahmen von einem Pre-Test und einem Post-Test erhoben. Die vollständigen Pre- und Post-Tests sind in Anhang B zu finden. Im *Pre-Test* wurden die folgenden Grundprofildaten erfragt: Geschlecht, Herkunftsland, Deutschkenntnisse sowie Dauer und Bereich der Übersetzungserfahrung.

Der *Post-Test* bestand aus den folgenden Teilen:

Teil 1: Fachliche Fragen zur Einstellung gegenüber der MÜ sowie zu Kenntnissen bzw. Erfahrung mit der KS. Diese Fragen wurden erst nach der Evaluation gestellt, damit die Teilnehmer keinen Hinweis im Vorfeld bekommen, worum es sich bei der Bewertung handelt.

Teil 2: Feedback zur Evaluation bzw. zur Testphase

Eine Analyse der Testdaten bzw. der Teilnehmerprofildaten und ihrer Feedbacks ist unter §5.2.5 aufgeführt.

4.5.5.4 Struktur der Ergebnisse der Humanevaluation

Im Rahmen der Humanevaluation erfolgte der Vergleich der Szenarien vor-KS vs. nach-KS im Hinblick auf die Stil- und Inhaltsqualität. Zudem konnten die Korrelationen zwischen den Fehlertypen und der Stil- und Inhaltsqualität untersucht sowie die Stil- und Inhaltsqualität der Annotationsgruppen analysiert werden. Konkret konnte anhand der Humanevaluation Folgendes ermittelt bzw. realisiert werden:

(1) Vergleich der Stil- und Inhaltsqualität-Scores der MÜ vor vs. nach der Anwendung der einzelnen KS-Regeln. Hierfür wurde der Signifikanztest Wilcoxon verwendet, da nicht alle Qualitätswerte normalverteilt waren.

(2) Berechnung und Analyse der Korrelation zwischen den Fehlertypen und der Stil- und Inhaltsqualität. Die Korrelation wurde mithilfe des Spearman-Korrelationstests berechnet, da die Differenz der Fehleranzahl der einzelnen Fehlertypen ordinal war. Zudem setzt Spearman keine Anforderung an die Verteilung und die Linearität voraus.

(3) Untersuchung der einzelnen Kriterien der Stil- und Inhaltsqualität

(4) Vergleich der Stil- und Inhaltsqualität auf Annotationsgruppenebene. Zur Messung der Signifikanz wurde der Wilcoxon-Test verwendet, da nicht alle Qualitätswerte normalverteilt waren.

(5) Erhalt der Alternativübersetzungen, die für die Ermittlung der MÜ-Qualität mithilfe der automatischen Evaluationsmetriken erforderlich sind.

4.5.6 Design der automatischen Evaluation

Um die Qualität der MÜ vor und nach dem Einsatz der einzelnen KS-Regeln aus einer anderen Perspektive zu beurteilen, wurde eine automatische Evaluation durchgeführt. Hierfür wurden die im Rahmen der Humanevaluation vorgeschlagenen Alternativübersetzungen herangezogen und als Referenzübersetzung für die Evaluation mithilfe von zwei automatischen Evaluationsmetriken (AEMs), nämlich TERbase und hLEPOR, eingesetzt. Durch die Triangulation der Ergebnisse der Humanevaluation mit denen der automatischen Humanevaluation konnte die Korrelation zwischen den Scores der AEMs und der Stil- und Inhaltsqualität genauer untersucht werden. Anhand der triangulierten Ergebnisse konnte die Richtung der Qualitätsveränderung durch eine weitere Evaluationsmethode belegt werden. In den nachstehenden Abschnitten wird das Testdesign näher erläutert sowie die Auswahl der beiden genannten Evaluationsmetriken begründet.

4.5.6.1 Testdesign

Wie das Testdesign der Humanevaluation zeigte, bestand ein Teil der Testaufgaben darin, ein „Light Post-Editing“ durchzuführen. Ein „Light Post-Editing“ zielt darauf ab, den Inhalt in einer verständlichen und genauen Weise bereitzustellen und nur wesentliche Korrekturen vorzunehmen (vgl. O’Brien 2010b). Kritisiert der Teilnehmer nur einen Teil oder ein Wort innerhalb der Übersetzung, so sollte

er diesen Teil korrigieren oder eine Verbesserung dafür vorschlagen. Sollte seine Korrektur bzw. Verbesserung mehrere Stellen in der MÜ betreffen oder den Rest der MÜ grammatisch oder stilistisch beeinflussen, wurde der Teilnehmer aufgefordert, eine Alternativübersetzung für den gesamten Satz anzugeben. Im Folgenden der entsprechende Teil der Testanleitung:

> When is it necessary to enter an alternative translation for the whole sentence at the bottom?
>
> - If you just recommend replacing a certain part in the translation and agree with the rest of the translation, you do not need to enter an alternative translation.
> - If your suggested modification (e.g. in the main clause) requires a further modification (e.g. in the subordinate clause), an alternative translation for the whole sentence is needed; otherwise, the translation might not sound idiomatic.

Dementsprechend standen für jeden Ausgangssatz bis zu acht Alternativübersetzungen (eine Alternativübersetzung pro Teilnehmer im Fall, dass alle Teilnehmer die MÜ kritisierten) zur Verfügung. Diese Alternativübersetzungen agierten als Referenzübersetzungen für die AEMs, vergleichbar mit der Vorgehensweise von Snover u. a. (2006).

Vergleicht man posteditierte MÜ mit Humanreferenzübersetzungen, die beide korrekt sind, sind die posteditierten MÜ in der Regel den MÜ ähnlicher als die Humanreferenzübersetzung (vgl. Denkowski & Lavie 2012). Gleichzeitig besteht eines der bekanntesten Probleme der AEMs darin, „that metrics are good at detecting similar translations, but poor at evaluating sentences with different structure and lexical choice“ (ebd.: 6). Vor diesem Hintergrund könnte die Verwendung von Humanreferenzübersetzungen mit schlechten AEM-Scores – nur aufgrund der Unterschiede zwischen der MÜ und der Humanreferenzübersetzung – verbunden sein, obwohl die MÜ korrekt ist. Mit der Verwendung der posteditierten MÜ konnte die Studie dieses Risiko umgehen.

Im Folgenden wird die Vorgehensweise bei der Auswahl der Referenzübersetzungen sowie die Vergleichsbasis der Szenarien vor-KS vs. nach-KS aus Sicht der automatischen Evaluation erklärt.

4.5.6.2 Auswahl der Referenzübersetzungen

Aus den verfügbaren Alternativübersetzungen aller Teilnehmer wurden Übersetzungen zweier Teilnehmer pro Ausgangssatzpaar (vor und nach KS) ausgewählt, die für die AEMs als Referenzübersetzungen fungierten. Wie Denkowski & Lavie (2012) zeigten, kann der MÜ-Akzeptanzgrad des Humanübersetzers stark variieren. Um unterschiedliche PE-Möglichkeiten zu berücksichtigen, wurde die automatische Evaluation auf Basis von Referenzübersetzungen von *zwei* Teilnehmern durchgeführt. In einer weiteren Studie mit dem Titel „Potential and limits of using post-edits as reference translations for MT evaluation" fanden Popović u. a. (2016) heraus, dass posteditierte MÜ-Outputs „definitely useful as reference translations" sind. Sie empfahlen gleichzeitig die Verwendung des PE-Outputs einer qualitativen MÜ eines unabhängigen Systems (d. h. PE-Output eines anderen Systems) sowie die Verwendung mehrerer Referenzen (ebd.). Mehr als zwei Referenzübersetzungen pro Ausgangssatz waren nicht immer verfügbar, da z. B. korrekte MÜ kein PE erfordern. Es wurde außerdem, wie von Popovic et al. (ebd.) empfohlen, darauf geachtet, dass das PE eines anderen Systems verwendet wurde (z. B. wurde bei der automatischen Evaluation von Systran das PE von Lucy verwendet). Zudem wurde bei der Auswahl der Referenzübersetzungen berücksichtigt, dass möglichst von allen Teilnehmern gleichermaßen Übersetzungen vertreten sind. Zunächst wurden die Teilnehmer chronologisch nach Abgabedatum des letzten Tests aufgelistet. Die Alternativübersetzung des ersten Teilnehmers wurde, falls vorhanden, ausgewählt. Sollte der erste Teilnehmer keine Alternativübersetzung vorgeschlagen haben, wurde die Übersetzung des zweiten Teilnehmers genommen. Bei dem nächsten Satz wurde die Übersetzung des dritten Teilnehmers ausgewählt usw.

In den Testanweisungen erhielten die Teilnehmer die allgemeine Hintergrundinformation, dass es sich bei der Übersetzung um Benutzerhandbücher und Gebrauchsanweisungen handelt. Es wurden keine weiteren Angaben zum genauen Kontext der einzelnen Testsätze zur Verfügung gestellt. Ziel hierbei war, dass die Teilnehmer aus einer vergleichbaren Ausgangssituation wie die MÜ-Systeme heraus kontextlos handeln und auf dieser Basis die MÜ-Sätze bewerten und ggf. Alternativübersetzungen vorschlagen. Nach den Testanleitungen sollten die Teilnehmer einen Kommentar hinterlassen, wenn sie bei einem bestimmten Satz Kontextinformation benötigten, um die MÜ bewerten oder eine Alternativübersetzung vorschlagen zu können. Im Endeffekt war die Anzahl der Kommentare zu fehlenden Kontextinformationen sehr gering und die Kommentare flossen bei der qualitativen Auswertung mit ein. Gleichzeitig wurden die Alternativübersetzungen mit diesen Kommentaren als Referenzübersetzungen in den AEMs ver-

wendet.

4.5.6.3 Auswahl der automatischen Evaluationsmetriken

Wie im Kapitel MÜ (siehe §3.4.3.2 „Automatische Evaluation“) dargestellt wurde, wurden seit den 90er Jahren zahlreiche AEMs entwickelt. In dieser Studie wurden die AEMs TERbase (Snover u. a. 2006) und hLEPOR (Han u. a. 2013) verwendet.

TER[57] (Snover u. a. 2006) zählt zu den populären AEMs zum Systemvergleich für europäische Sprachen (vgl. Doherty 2017). In TER haben alle Editierungsvorgänge (Edit operations) (Einfügen, Löschen, Ersetzen und Verschieben) einheitliche Editierungskosten (Edit cost) von eins. TERbase ist eine Variante der Metrik TER (Gonzàlez & Giménez 2014: 19). Anders als TER führt TERbase kein „Synonym Match“ (vgl. ebd.). In der Studie führte die Verwendung von Synonymen bei einer KS-Regel wie z. B. „Eindeutige pronominale Bezüge verwenden“ zu einer lexikalischen Inkonsistenz (siehe Tabelle 5.67). Entsprechend war es erforderlich, dass ein Punktabzug bei der Verwendung von Synonymen erfolgte.[58] Aus diesem Grund wurde TERbase verwendet. Die Qualität-Scores von TER bzw. TERbase bewegen sich zwischen − 1 (niedrigster Wert) und 0 (höchster Wert); somit arbeitet TERbase mit negativen Werten. Die Evaluation mithilfe von TERbase wurde mit dem online verfügbaren Tool Asiya[59] (Gonzàlez & Giménez 2014) durchgeführt. Das Tool ist benutzerfreundlich und kostenfrei online zugänglich.

Eine weitere sehr verbreitete AEM für die MÜ-Evaluation zwischen europäischen Sprachen ist BLEU[60] (Papineni u. a. 2002). Trotz ihrer weit verbreiteten Nutzung konnte BLEU in der Studie nicht verwendet werden. Die Evaluation im Rahmen dieser Studie erfolgt auf Satzebene, während BLEU auf Dokumentebene funktioniert (ebd.). Zudem zeigen Callison-Burch u. a. (2009), dass Metriken der neuen Generation BLEU hinsichtlich der Korrelation mit der Humanevaluation übertreffen.

Eine hohe Korrelation mit der Humanevaluation ist ein Hauptkriterium bei der Bewertung der Effektivität von AEMs, wodurch AEMs die kostenintensive Humanevaluation vermehrt ersetzen können (vgl. Lin & Och 2004). Vor diesem Hintergrund war es erforderlich, eine zweite AEM der neuen Generation zu finden, deren Scores mit den Bewertungen im Rahmen von Humanevaluationen stark korrelieren. Im WMT Workshop 2011[61] zeigte die Spearman Korrelation

[57] TER steht für Translation Edit Rate (Snover u. a. 2006).

[58] Ebenfalls wurde Meteor (Banerjee & Lavie 2005) aus demselben Grund, nämlich dass kein Punktabzug bei der Verwendung von Synonymen erfolgt, ausgeschlossen.

[59] Asiya ist online zugänglich unter: http://asiya.cs.upc.edu/demo/asiya_online.php

[60] BLEU steht für BiLingual Evaluation Understudy (Papineni u. a. 2002).

[61] http://www.statmt.org/wmt11

zwischen den Bewertungen der Humanevaluation und Scores der AEM hLEPOR die höchste Korrelation im Vergleich zu denen der AEMs MPF, ROSE, METEOR, BLEU und TER – sowohl sprachübergreifend (Korrelationskoeffizient von 0,83) als auch für das Sprachenpaar der vorliegenden Studie DE>EN (Korrelationskoeffizient 0,86) (Han u. a. 2013). Das Berechnungsmodell von hLEPOR[62] basiert auf einem harmonischen Mittelwert von drei Faktoren: einer „enhanced length penalty", einer „N-gram position difference penalty" sowie „precision and recall" (Han u. a. 2013). Die Qualität-Scores von hLEPOR bewegen sich zwischen 0 (niedrigster Wert) und 1 (höchster Wert); somit arbeitet hLEPOR mit positiven Werten. Dank des Entwickler-Teams von hLEPOR wurde ein einfach zu bedienendes Tool mit einem für DE-EN getunten Skript zu Forschungszwecken im Rahmen der Studie kostenfrei zur Verfügung gestellt.

Aus den obengenannten Motiven sowie Kosten- und Effizienzgründen wurde die automatische Evaluation nur mithilfe der AEMs TERbase und hLEPOR durchgeführt. Weitere AEMs, die über vergleichbare Merkmale verfügen, sind somit nicht auszuschließen und können in zukünftigen Analysen zum Vergleich herangezogen werden.

4.5.6.4 Basis des Vergleichs vor-KS vs. nach-KS zur Ermittlung des KS-Einflusses

Der Qualitätsvergleich der beiden Szenarien vor-KS vs. nach-KS erfolgte in der automatischen Evaluation mit den beiden AEMs (hLEPOR und TERbase) wie folgt:

(1) Für jedes Szenario wurden die AEM-Scores anhand zweier Referenzübersetzungen berechnet. Beispielsweise wurde für Satz 1 im Szenario vor-KS der hLEPOR-Score auf Basis der Referenzübersetzungen 1 und 2 (Ref.1 und Ref.2) berechnet. Dies ergab die Scores hLEPOR-Ref.1 und hLEPOR-Ref.2).

(2) Der Mittelwert der Scores aus den beiden Referenzübersetzungen wurde berechnet (d. h. im Beispiel der Mittelwert der Scores hLEPOR-Ref.1 und hLEPOR-Ref.2).

(3) Nach derselben Vorgehensweise wurde der Mittelwert für das Szenario nach-KS berechnet.

[62] hLEPOR steht für Harmonic mean of enhanced Length Penalty, Precision, n-gram Position difference Penalty and Recall (Han u. a. 2013).

(4) Der Einfluss der KS-Regel wurde auf Basis der Differenz „Mittelwert nach-KS“ *minus* „Mittelwert vor-KS“ gemessen. Eine positive Differenz deutet auf eine Verbesserung des AEM-Scores hin, und umgekehrt weist eine negative Differenz auf eine Verschlechterung des AEM-Scores hin.

Obwohl mehrere AEMs die Möglichkeit bieten, parallel diverse Referenzübersetzungen zu verwenden (vgl. Han 2018) – wobei nur die Referenzübersetzung mit dem höheren Match für die Score-Berechnung herangezogen wird –, wurde der Metrik-Score in dieser Studie in zwei Durchgängen jeweils nach einer Referenzübersetzung gesondert berechnet. Ziel dabei war es, die Alternativübersetzungen von zwei Teilnehmern zu berücksichtigen, anstatt die Evaluation auf Basis einer Referenzübersetzung einzuschränken.

Es bestand ferner die Möglichkeit, dass die Teilnehmer Stellen außerhalb der KS-Stelle posteditieren. Daher wurde (1) darauf geachtet, dass die Referenzübersetzungen der vor- und nach-Szenarien vom selben Teilnehmer stammen. Sollte der Teilnehmer eine Stelle außerhalb der KS-Stelle posteditieren, würde diese Stelle in den Referenzübersetzungen (vor- und nach-KS) wiederholt. (2) Die Ermittlung des KS-Einflusses erfolgte auf Basis der Differenz ‚AEM-Score nach-KS‘ *minus* ‚AEM-Score vor-KS‘, damit der im Score enthaltene Anteil von potenziellen Edits außerhalb der KS-Stelle herausgefiltert werden kann. Tabelle 4.13 veranschaulicht die Problematik des PE einer Stelle außerhalb der KS-Stelle und zeigt wie sie gelöst wurde:

In Tabelle 4.13 ging es um die KS-Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“. Dementsprechend besteht die KS-Stelle aus dem Wort *Raumdruck* ohne Anführungszeichen im Szenario vor-KS bzw. mit Anführungszeichen im Szenario nach-KS (fett hervorgehoben). Die Edits in den Referenzübersetzungen wurden innerhalb der KS-Stelle unterstrichen und außerhalb der KS-Stelle farblich hervorgehoben. Wie unter §4.5.3.1 ([9] Aufbereitung der Zielsätze für die Humanevaluation) erläutert wurde, wurden die MÜ-Sätze außerhalb der KS-Stelle vor und nach dem Einsatz der KS-Regeln vereinheitlicht. Dies ließ die Teilnehmer-Edits außerhalb der KS-Stelle – falls Edits erforderlich waren – ebenfalls einheitlich. Der erste Teilnehmer (Ref.1) editierte in seiner Alternativübersetzung nur die KS-Stelle. Der zweite Teilnehmer (Ref.2) hingegen editierte sowohl die KS-Stelle als auch eine Stelle außerhalb der KS-Stelle (vgl. ‚can be activated‘).

Der Einfluss der KS errechnete sich in dem Beispiel wie folgt = [(0,9682 + 0,6947) / 2] − [(0,8794 + 0,5953) / 2] = 0,0941. Durch die Subtraktion Score-nach-KS *minus* Score-vor-KS wurde bei der zweiten Referenzübersetzung (Ref.2) der Effekt der außerhalb der KS-Stelle editierten Stelle (‚can be activated‘) herausgefiltert,

Tabelle 4.13: Beispiel 9

		hLEPOR
Vor KS	Die zwei aktivierbaren Raumdruck-Sollwerte sind im Menü **Raumdruck** definiert.	
MÜ vor KS	The two activatable room pressure setpoints are defined in the **area pressure** menu.	
Ref.1	The two activatable room pressure setpoints are defined in the "**Room pressure**" menu.	0,8794
Ref.2	The two room pressure target values that can be activated are defined in the "**room pressure**" menu.	0,5953
		hLEPOR
Nach KS	Die zwei aktivierbaren Raumdruck-Sollwerte sind im Menü "**Raumdruck**" definiert.	
MÜ nach KS	The two activatable room pressure setpoints are defined in the "**room pressure**" menu.	
Ref.1	The two activatable room pressure setpoints are defined in the "**Room pressure**" menu.	0,9682
Ref.2	The two room pressure target values that can be activated are defined in the "**room pressure**" menu.	0,6947

KS-Stelle ist **fett** dargestellt; Edits innerhalb der KS-Stelle sind **unterstrichen**; Edits außerhalb der KS-Stelle sind hervorgehoben.

denn diese Stelle wiederholte sich vor- und nach-KS. Der berechnete positive Differenzwert (0,0941) deutet auf einen positiven Einfluss der KS-Regel hin.

4.5.6.5 Struktur der Ergebnisse der automatischen Evaluation

Im Rahmen der automatischen Evaluation erfolgte der Vergleich der Szenarien vor-KS vs. nach-KS mithilfe der Qualitätsmetriken TERbase und hLEPOR. Zudem wurden die Korrelationen zwischen den Scores der Qualitätsmetriken und der Stil- und Inhaltsqualität untersucht. Konkret konnte anhand der automatischen Evaluation Folgendes ermittelt bzw. realisiert werden:

(1) Vergleich der Scores der Qualitätsmetriken vor vs. nach der Verwendung der einzelnen KS-Regeln. Hierbei wurde der Signifikanztest Wilcoxon verwendet, da nicht alle Qualitätswerte normalverteilt waren.

(2) Berechnung und Analyse der Korrelation zwischen den Scores der Qualitätsmetriken und der Stil- und Inhaltsqualität. Die Korrelation wurde mithilfe des Spearman-Korrelationstests berechnet, da nicht alle Qualitätswerte normalverteilt waren. Zudem setzt Spearman keine Anforderung an die Verteilung und die Linearität voraus.

4.6 Fazit

Mithilfe eines dreiphasigen Mixed-Methods-Triangulationsansatzes wird der Einfluss einzelner Regeln der Kontrollierten Sprache auf die Qualität des MÜ-Outputs unterschiedlicher MÜ-Systeme untersucht und verglichen. Die implementierten Methoden sind: Fehlerannotation, Humanevaluation und automatische Evaluation. Jede Methode liefert Daten, die für die Analyse in der darauffolgenden Methode erforderlich sind. Auf diese Weise wurde angestrebt, das Ziel der Studie über drei Phasen systematisch zu realisieren und dabei sämtliche Herausforderungen der KS-Untersuchung auf Regelebene zu überwinden. Ferner wurden die Ergebnisse der Analysen trianguliert, um diverse Fragestellungen beantworten zu können. Nicht nur eine methodenexterne Triangulation der drei genannten Methoden sah die angewandte Methodik vor, sondern auch eine methodeninterne Triangulation beim Design der Humanevaluation. Im Testdesign der Humanevaluation waren die Qualitätsdefinitionen in die anzukreuzenden Qualitätskriterien integriert. Somit hatten alle Teilnehmer eine direkte und einheitliche Basis für die Vergabe der Qualität-Scores. Des Weiteren mussten die Teilnehmer vor der Vergabe der Qualität-Scores die angekreuzten Qualitätskriterien

kommentieren bzw. posteditieren. Diese Triangulation fördert die interne Konsistenz, optimiert die Zuverlässigkeit und ermöglicht eine genaue Interpretation der Daten. Ferner ist das Testdesign replizierbar. In diesem Kapitel wurde die Methodologie der Studie unter Einbeziehung der Operationalisierung und der Validität präsentiert. Zur Entwicklung des Studiendesigns wurden die theoretische Basis sämtlicher relevanten Aspekte sowie die Ergebnisse vorheriger Studien reflektierend diskutiert. Das daraus resultierende Design wurde bei der Datenerfassung und -analyse umgesetzt. Die Ergebnisse sind im folgenden Kapitel aufgeführt.

5 Quantitative und qualitative Analyse der Ergebnisse

5.1 Einleitung

In diesem Kapitel werden die Ergebnisse quantitativ dargestellt und anhand von Beispielen aus dem Datensatz qualitativ erläutert. Das Kapitel besteht aus vier Unterkapiteln: Im ersten Unterkapitel werden die Ergebnisse einer allgemeinen Analyse des Datensatzes sowie der Teilnehmerprofildaten und -feedbacks zur Evaluation präsentiert. Die drei weiteren Unterkapitel liefern die Ergebnisse auf den vier Analyseebenen: der Sprachenpaarebene (unter §5.3), der Regelebene sowie der Regel- und MÜ-Systemebene (unter §5.4) und der MÜ-Systemebene (§5.5).

5.2 Allgemeine Analyse

In diesem Unterkapitel werden die Ergebnisse einer allgemeinen Analyse dargestellt. Diese Analyse umfasst einen Überblick über den untersuchten Datensatz, die Methode der Festlegung der Teilnehmeranzahl bei der Humanevaluation, eine Darstellung der Interrater- und Intrarater-Agreements sowie eine genaue Betrachtung der Teilnehmerprofildaten, der Teilnehmer-Feedbacks zur Evaluation und ihrer Implikationen.

5.2.1 Überblick über den Datensatz

Zunächst liefert Tabelle 5.1 einen Überblick über den Datensatz, der im Rahmen der ersten Analyse (Fehlerannotation) und der zweiten Analyse (Humanevaluation) untersucht wurde. Insgesamt wurden 2.160 Sätze annotiert. Daraus wurden 1.100 von den teilnehmenden Übersetzern bewertet. Darüber hinaus gab es insgesamt 545 MÜ-Sätze, die bei mehreren MÜ-Systemen identisch waren (d. h. die Ausgangssätze wurden von verschiedenen Systemen identisch übersetzt). Für jeden Ausgangssatz wurde nur eine Instanz der identischen MÜ-Sätze in der

Humanevaluation bewertet, somit waren 95 der 545 Instanzen in den 1.100 humanevaluierten MÜ-Sätzen enthalten. Alle weiteren identischen wiederholten Instanzen (450 von 545) erhielten denselben Score der bewerteten Instanzen. Basis der präsentierten statistischen Ergebnisse ist die Summe der humanevaluierten (1.100) und der wiederholten (450) MÜ-Sätze (d. h. insgesamt 1.550 MÜ-Sätze). Die restlichen 610 Sätze vom Gesamtdatensatz wurden ausgeschlossen. Die Ausschlusskriterien sind unter §4.5.3.1 [8] und [9] dargestellt.

Tabelle 5.1: Überblick über den Datensatz

Anzahl der annotierten MÜ (Gesamtdatensatz):	2.160	(100 %)
Anzahl der humanevaluierten MÜ:	1.100	(51 %)
Anzahl der wiederholten MÜ:	450	(21 %)
Anzahl der ausgeschlossenen MÜ:	610	(28 %)

5.2.2 Entwicklung des Mittelwerts der Qualität mit der Zunahme der Teilnehmeranzahl

Zur Festlegung der Anzahl der teilnehmenden Bewerter bei der Humanevaluation wurde zunächst mit fünf Teilnehmern getestet und die Anzahl der Teilnehmer sukzessive erhöht bis die akkumulierten Mittelwerte der Stil- und Inhaltsqualität sich stabilisierten (Genaueres zur Teilnehmeranzahl sowie Qualitätsdefinition unter §4.5.5.3 bzw. §4.5.5.1). Wie Abbildung 5.1 zeigt, begannen sich ab dem 6. Teilnehmer und bis zu dem 8. Teilnehmer die akkumulierten Mittelwerte der Stil- und Inhaltsqualität zu stabilisieren und veränderten sich kaum. Entsprechend wurde die Teilnehmeranzahl nicht weiter erhöht.

5.2.3 Interrater-Agreement

Zur Messung des Interrater-Agreements (Interrater-Reliabilität) in der Humanevaluation wurde eine Reliabilitätsanalyse durchgeführt, in der die „Intra-Klassen-Korrelation“ (ICC) berechnet wurde. Die ICC wird zur Berechnung des Agreements zwischen mehr als zwei Beobachtern und bei metrischen Variablen (hier eine 1-5-Likertskala) verwendet (vgl. Landis & Koch 1977).

Wie Tabelle 5.2 zeigt, ist das Interrater-Agreement bei allen Qualitätsmessungen (Stilqualität „SQ“ und Inhaltsqualität „CQ“ vor und nach der Anwendung der KS) größer als 0,8, mit Ausnahme der Stilqualität nach KS, die knapp unter 0,8 lag. Nach Landis & Koch (1977) deutet dieses Ergebnis auf eine „substantial“ bis

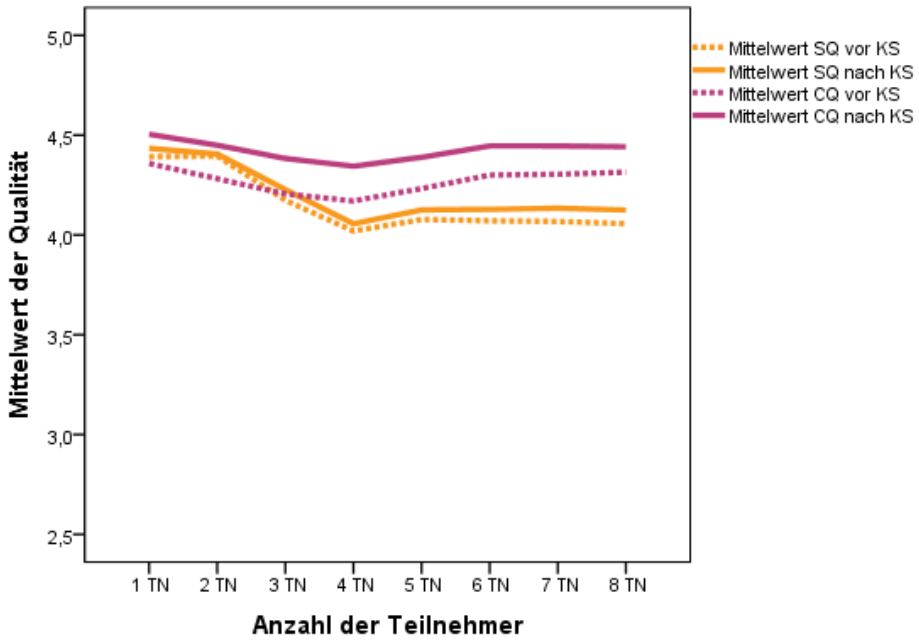


Abbildung 5.1: Entwicklung des Mittelwerts der Qualität mit der Zunahme der Teilnehmeranzahl

Tabelle 5.2: Interrater-Agreement

Qualität mit 8 TN	ICC
SQ vor KS	0,816
SQ nach KS	0,763
CQ vor KS	0,891
CQ nach KS	0,890

0 = poor 0 – 0,2 slight 0,2 – 0.4 fair
0,4 – 0,6 moderate 0,6 – 0,8 substantial 0,8 – 1 (almost) perfect

Interpretation der Koeffizienten als Grad der Übereinstimmung nach Landis & Koch (1977)

„(almost) perfect“ Übereinstimmung der Bewertungen der acht Teilnehmer hin (Tabelle 5.2).

5.2.4 Intrarater-Agreement

Der Datensatz beinhaltete 130 Übersetzungen (von den 2.160 Übersetzungen), die vor und nach der Anwendung der jeweiligen KS-Regel identisch waren (Tabelle 5.3).

Tabelle 5.3: Beispiel einer identischen MÜ vor und nach KS aus der Regel „Überflüssige Präfixe vermeiden“

Vor-KS	**Schicken** Sie das Gerät zusammen mit dem Original-Kaufbeleg an nachstehende Adresse **zu**.
Nach-KS	**Schicken** Sie das Gerät zusammen mit dem Original-Kaufbeleg an nachstehende Adresse.
Identische MÜ	**Send** the device together with the original receipt to the following address.

Von den 130 Übersetzungen wurden 102 (78,4 %) in der Humanevaluation bewertet (siehe Tabelle 5.4). Dies bot eine wertvolle Gelegenheit dafür, das Intrarater-Agreement (Intrarater-Reliabilität) zu messen. Tabelle 5.4 fasst die Anzahl der identischen Übersetzungen sowie der daraus in der Humanevaluation getesteten Übersetzungen bei jeder KS-Regel zusammen:

Zur Messung des Intrarater-Agreements (Intrarater-Reliabilität) in der Humanevaluation wurde eine Reliabilitätsanalyse durchgeführt, in der die „Intra-Klassen-Korrelation“ (ICC) berechnet wurde. Die ICC wird zur Berechnung des Agreements zwischen mehr als zwei Beobachtern und bei metrischen Variablen (hier eine 1-5-Likertskala) verwendet (vgl. Landis & Koch 1977).

Nach den Ergebnissen (Tabelle 5.5) lag das Intrarater-Agreement bei der Inhaltsqualität bei fünf der acht Teilnehmer bei mindestens 0,8 („(almost) perfect“ Übereinstimmung nach Landis & Koch (1977)) und bei drei Teilnehmern zwischen 0,6 und 0,8 („substantial“ Übereinstimmung, ebd.). Auf der anderen Seite lag bei der Stilqualität das Intrarater-Agreement nur bei zwei der acht Teilnehmer über 0,8 („(almost) perfect“ Übereinstimmung, ebd.), bei drei Teilnehmern zwischen 0,7 und 0,8 („substantial“ Übereinstimmung, ebd.), bei zwei Teilnehmern bei ca. 0,5 („moderate“ Übereinstimmung, ebd.) und beim letzten Teilnehmer bei ca. 0,3

[1]Die Namen der Teilnehmer wurden aus Datenschutzgründen abgekürzt.

Tabelle 5.4: Anzahl der identischen MÜ und der davon in der Humanevaluation getesteten MÜ bei jeder Regel

	Identische MÜ	Getestete identische MÜ	%
per	2	2	2/2, 100 %
pak	4	4	4/4, 100 %
nsp	3	3	3/3, 100 %
kos	39	31	31/39, 79 %
wte	6	5	5/6, 83 %
pas	0	0	0
FVG	14	10	10/14, 71 %
Prä	61	46	46/61, 75 %
anz	1	1	1/1, 100 %
Summe	130	102	102/130, 78 %

Tabelle 5.5: Intrarater-Agreement

	ICC - SQ	ICC - CQ
Jam	0,744	0,923
Ryn	0,381	0,800
Ros	0,599	0,739
Ven	0,877	0,899
Tan	0,806	0,849
Cro	0,790	0,835
Stl	0,565	0,706
Len[1]	0,716	0,611

0 = poor 0 – 0,2 slight 0,2 – 0.4 fair
0,4 – 0,6 moderate 0,6 – 0,8 substantial 0,8 – 1 (almost) perfect

Interpretation der Koeffizienten als Grad der Übereinstimmung nach Landis & Koch (1977)

(„fair" Übereinstimmung, ebd.) (Tabelle 5.5). Eine mögliche Interpretation der Übereinstimmungsunterschiede zwischen der Inhalts- und Stilqualität ist, dass die Bewertung der Inhaltsqualität hinsichtlich der Genauigkeit und der Verständlichkeit in der Regel eine unumstritten eindeutigere Aufgabe im Vergleich zu der Bewertung der Stilqualität ist. Vor diesem Hintergrund werden in einigen Post-Editing-Leitlinien inhaltliche Fehler behoben und stilistische Probleme ignoriert (O'Brien u. a. 2009; O'Brien 2010b), da in der Regel mehrere stilistisch akzeptable Formulierungen denkbar sind.

5.2.5 Analyse der Teilnehmerprofildaten und -feedbacks zur Evaluation

Im Rahmen der Humanevaluation wurden die Profildaten anhand des Pretests und Posttests (siehe Anhang B) erhoben. Im *Pretest* wurden die Grunddaten zum Geschlecht, Herkunftsland, Deutschkenntnisse und Übersetzungserfahrung erfragt.

Wie Tabelle 5.6 (Pretest) zeigt, waren die Teilnehmerprofile bezüglich des Geschlechts ausgewogen. Sie stammen aus zwei englischsprachigen Ländern, beherrschen die deutsche Sprache und verfügen über Übersetzungserfahrung. Ferner wurde die Homogenität der Teilnehmer bereits bei den Teilnahmekriterien (Besitz eines Bachelor-Abschlusses in Translation und das Studium im letzten Semester des Master-Abschlusses Translation) sichergestellt.

Der *Posttest* bestand aus zwei Teilen: Teil 1 *Fachliche Fragen* und Teil 2 *Feedback zur Evaluation bzw. zur Testphase.* Die *Fachlichen Fragen* (Teil 1) wurden erst nach der Evaluation gestellt, damit die Teilnehmer im Vorfeld keinen Hinweis bekommen, worum es sich bei der Bewertung handelt. Dieser Teil bestand aus zwei Fragen:

1. **Einstellung zur MÜ:** Die Einstellung zur MÜ wurde durch die Verwendung von MÜ-Systemen erfragt. Die Frage *Wie gehst Du in der Regel vor, wenn Du einen technischen Text übersetzen möchtest? bzw. wie würdest Du vorgehen?* zeigte Folgendes:

   Vier der acht Teilnehmer verwenden ein Übersetzungsprogram mit einem Zugriff auf ein MÜ-System zum Übersetzen und anschließend posteditieren sie ihre maschinellen Übersetzungen; zwei Teilnehmer konsultieren ein MÜ-System nur bei Bedarf; zwei Teilnehmer verwenden gar keine MÜ-Systeme, da sie sich mit der Übersetzung literarischer Texte beschäftigen. Auf Basis dieses Ergebnisses wird in der Studie von einer überwiegend positiven Einstellung zur MÜ ausgegangen.

Tabelle 5.6: Profildaten der Teilnehmer

Pretest	Profildaten der Teilnehmer
Geschlecht	Vier Studentinnen und vier Studenten
Herkunftsland	Zwei Teilnehmer aus Großbritannien; sechs aus den USA
Deutschkenntnisse	Zwei Teilnehmer sind zweisprachig (Englisch/Deutsch) aufgewachsen, wobei sie Englisch als ihre dominante Sprache einschätzen. Drei Teilnehmer erwarben Kenntnisse der deutschen Sprache durch den Besuch einer deutschen Schule im Herkunftsland. Drei Teilnehmer erlernten die deutsche Sprache im Rahmen von Deutschsprachkursen sowie durch einen vierjährigen Aufenthalt in Deutschland.
Dauer der Übersetzungserfahrung	Fünf Teilnehmer haben ein Jahr Übersetzungserfahrung. Zwei Teilnehmer haben zwei Jahre Übersetzungserfahrung. Ein Teilnehmer hat vier Jahre Teilzeit Übersetzungserfahrung.

2. **Kenntnisse bzw. Erfahrung mit der Kontrollierten Sprache** wurden wie folgt erfragt: *Hast Du Erfahrung mit der Kontrollierten Sprache (Controlled Language)?*

   Die Antworten der Teilnehmer zeigten, dass sechs Teilnehmern das Thema KS durch das Studium bekannt ist; zwei Teilnehmer haben zudem Übersetzungserfahrung mit der KS; zwei Teilnehmer hatten bislang gar keinen Kontakt mit dem Thema. Die Auswahlmöglichkeiten „Ich beachte im Allgemeinen bei meiner technischen Übersetzung die Regeln der Kontrollierten Sprache“ und „Obwohl die Regeln der Kontrollierten Sprache mir bekannt sind, bevorzuge ich danach nicht zu übersetzen“ wurden von keinem Teilnehmer angekreuzt. Da die Mehrheit der Teilnehmer keine berufliche Erfahrung mit der KS hatte, war ihre Einstellung zur KS noch nicht gefestigt (weder positiv noch negativ). Auf Basis dieses Ergebnisses wird von einer neutralen Einstellung zur KS ausgegangen.

Im *Feedback zur Evaluation bzw. zur Testphase* (Teil 2) wurden sechs Fragen über den Umfang, Aufbau und Inhalt der Evaluation gestellt und ihre Implikationen bei der Auswertung berücksichtigt. Dieses Feedback kann außerdem als Erkenntnis bei zukünftigen Untersuchungen dienen. Im Folgenden wurden die Antworten der Teilnehmer zusammengefasst und analysiert:

1. Fandst du die Evaluation lang / umfangreich?
   Sechs der Teilnehmer fanden den Umfang der Evaluation eher groß, dennoch aufgrund der Zeit- und Ortsflexibilität machbar. Kommentare in diesem Zusammenhang waren: „Es war ziemlich viel, aber ich würde nicht sagen, dass es zu viel war, weil ich genug Zeit hatte. Wenn ich das alles in 2 Wochen hätte schaffen müssen, wäre das zu viel gewesen" und „die Tests an sich waren meiner Meinung nach nicht unbedingt zu lang oder zu viel, ich hatte meinen Laptop dabei und es war möglich jeden Tag eine Stunde zu finden, in der ich in Ruhe arbeiten konnte." Die zwei weiteren Teilnehmer fanden den Umfang der Evaluation groß und gaben an, dass sie wenig Zeit hatten, vor allem weil sie während der Testphase ihre Masterarbeit schrieben.

   Bedeutung bzw. Implikation: Auf Basis dieses Feedbacks können der Evaluationsumfang und die dafür vorgesehene Zeitspanne in Kombination mit der Ortsflexibilität als akzeptabel beurteilt werden. Die Gewährung einer zeitlichen und örtlichen Freiheit und Flexibilität ist förderlich und sehr empfehlenswert.

2. Fandst du die Evaluation interessant / langweilig? Falls langweilig, wann hat die Langweile eingesetzt, nach dem 10., 20., …Test?
   Die meisten Teilnehmer (6 von 8) gaben an, dass die Evaluation am Anfang interessant war, jedoch etwa ab dem 20. Test begann, langweilig zu werden. Als Hauptgrund hierfür wurde die Wiederholung der MÜ-Sätze genannt. Des Weiteren fanden sie die Evaluation *„ziemlich langweilig, aber durch die Arbeit ohne Zeitdruck ziemlich angenehm"*. Der achte Teilnehmer beschrieb die Evaluation als *„interessant"*.

   Bedeutung bzw. Implikation: Bei manchen Ausgangssätzen waren die MÜ-Outputs ähnlich. Die Evaluation der feinen Unterschiede in den MÜ-Sätzen war für das Ziel der Studie unerlässlich. Die MÜ-Sätze wurden auf 44 Tests aufgeteilt und jeder Übersetzer hatte die Möglichkeit min. 1 und max. 3 Tests pro Tag zu bewerten. Bei ähnlichen Untersuchungen könnte es ratsam sein, die Zeitspanne der Evaluation zu verlängern oder die Anzahl der

ähnlichen MÜ-Sätze auf mehrere Teilnehmer aufzuteilen, d. h. eine Erhöhung der Anzahl der Teilnehmer mit einer gleichzeitigen Reduzierung der ähnlichen MÜ-Sätze pro Teilnehmer. Jedoch war es der Forscherin ein Anliegen, alle Testsätze komplett von jedem Teilnehmer bewerten zu lassen, um einem potenziellen negativen Einfluss auf das Agreement vorzubeugen.

3. Fandst du den Aufbau der Evaluation verständlich / zu schwer?
   Vier Teilnehmer beschrieben den Aufbau der Evaluation als „verständlich". Die anderen vier Teilnehmer gaben an, dass es am Anfang schwierig gewesen sei, die Fehler den Qualitätskriterien zuzuordnen.

   Bedeutung bzw. Implikation: Einige Teilnehmer wendeten sich mit Rückfragen an den Tester. Damit alle Teilnehmer mit derselben Ausgangssituation bzw. unter denselben Voraussetzungen die Evaluation durchführen konnten, bat der Tester sie, die Testanweisungen genauer zu lesen und wies sie auf die darin enthaltenen Beispiele hin. Eine Intervention durch den Tester wurde vermieden, damit jeder Teilnehmer im Rahmen der vorgeschriebenen Anweisungen sein eigenes Evaluationsschema festlegt und danach bewertet.

4. Hast Du durch die Bewertung etwas Neues gelernt?
   Drei Teilnehmer betrachteten die Evaluation als eine Korrekturaufgabe und fanden die genaue Analyse bzw. Differenzierung der Fehler lehrreich bzw. informativ. Ein Teilnehmerkommentar lautete „es wäre spannender, wenn die Übersetzungen unterschiedlich wären". Ein anderer Teilnehmer fand die Verwendung von MS-Excel für diesen Zweck „hilfreich". Die weiteren vier Teilnehmer ließen diese Frage unbeantwortet.

   Bedeutung bzw. Implikation: Ziel dieser Frage war indirekt Näheres über das Interesse bzw. die Motivation der Teilnehmer zu erfahren. Ebenfalls hier wäre es überlegungswert, die Anzahl der ähnlichen MÜ-Sätze auf mehrere Teilnehmer aufzuteilen (d. h. eine Erhöhung der Anzahl der Teilnehmer mit einer gleichzeitigen Reduzierung der ähnlichen MÜ-Sätze pro Teilnehmer). Jede Entscheidung hat ihre Vor- und Nachteile, die abgewogen werden müssen, wie unter 2 angegeben.

5. War in den DE- oder EN-Sätzen irgendetwas auffällig?
   Zwei Teilnehmer gaben an, dass sie aufgrund der Übersetzungen davon ausgehen, dass es sich um eine MÜ-Analyse handle. Vier Teilnehmer fanden, dass sich die Fehler oft wiederholten. Zwei Teilnehmer beantworteten die Frage mit „nein".

   Bedeutung bzw. Implikation: Ziel dieser Frage war indirekt zu erfahren, ob die Teilnehmer die KS-Regeln entdecken konnten, was den Antworten der Teilnehmer nach nicht der Fall war und durch die Antworten auf Frage 6 bestätigt wurde.

6. Kannst Du im Nachhinein erraten, worum es geht?
   Drei Teilnehmer antworteten, dass es sich um einen Vergleich von MÜ-Systemen handle. Die weiteren Antworten lauteten „Korrekturlesen und Fehleridentifikation" und „Verständlichkeitsforschung". Auch diese Frage wurde nicht von allen Teilnehmern beantwortet.

   Bedeutung bzw. Implikation: Um auszuschließen, dass die Teilnehmer das Ziel der Studie durch die Sätze erraten konnten und es entsprechend bei der Evaluation gezielt berücksichtigt zu haben, wurde diese Frage direkt gestellt. Den Antworten der Teilnehmer zufolge konnte keiner das Thema der Studie richtig erraten, was insbesondere bei der Bewertung der Stilqualität vorteilhaft ist. Dies kann auf die große Anzahl der bewerteten Sätze sowie die hohe Randomisierung der Sätze der verschiedenen Regeln zurückgeführt werden.

Zu den obengenannten Implikationen kann anhand des Teilnehmerfeedbacks ein mit der Humanevaluation verbundenes Risiko nahezu ausgeschlossen werden: White (2003) thematisiert die Gefahr, dass die Bewerter "react differently to a translated expression if they (think they) know how it got that way. […] in MT, judges will be more forgiving of particular errors if they think their cause is a trivial bug (e.g., missing lexical item) rather than a serious problem (e.g., scope of modification)". Weder das erhaltene Feedback im Posttest noch die Kommentare der Teilnehmer bei der Evaluation geben einen Hinweis darauf, dass die Bewertung der Teilnehmer durch den potenziellen technischen Grund der Fehler beeinflusst wurde. Die Teilnehmer, die MÜ-Systeme verwendeten, sind in dem Fall Nutzer mit Übersetzungskompetenz, jedoch besitzen sie kein technisches Wissen über die genaue Systemfunktionsweise bzw. keinen technischen Hintergrund, wie ein Fehler entstehen kann.

5.3 Analyse auf Sprachenpaarebene (regel- und systemübergreifend)

Fokus der Studie ist zwar der Vergleich des MÜ-Outputs vor vs. nach der Anwendung der *einzelnen* KS-Regel, für einen Gesamtüberblick bietet aber der folgende Abschnitt zunächst ein allgemeines einführendes Bild für diesen Vergleich regel- und systemübergreifend. Die hierfür angewendeten Analysefaktoren sind unter §5.3.1 genau dargestellt, gefolgt von den Ergebnissen dieser Faktoren. Bezüge auf vorherige Studien werden nicht in diesem Unterkapitel, sondern im Rahmen der Diskussion in §6 vorgenommen.

5.3.1 Analysefaktoren

Der Vergleich des MÜ-Outputs vor vs. nach der Anwendung aller analysierten KS-Regeln regel- und systemübergreifend erfolgte nach den folgenden neun Analysefaktoren:

5.3.1.1 Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln

- Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?
- Variablen: Summe der Fehler und Mittelwert der Fehleranzahl (von allen Fehlertypen) innerhalb der KS-Stelle; Variablentyp: ordinal
- Statistische Auswertung: Deskriptive Statistiken auf Basis der Fehlerannotation; Abbildungen: Balken und Fehlerbalken
- Hypothesen:

  H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.

  H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.
- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Die Variablen sind ordinal.

5.3.1.2 Zweiter Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln außerhalb der KS-Stelle bei der Gruppe RR

- Fragestellung: Wie häufig wurde die KS-Stelle vor und nach der Anwendung der KS-Regeln korrekt übersetzt und stieg *gleichzeitig* die Fehleranzahl *außerhalb* der KS-Stelle nach der Anwendung der KS-Regeln?
- Variablen: Differenzen in der Fehleranzahl bei der Annotationsgruppe RR (RR: MÜ innerhalb der KS-Stelle ist vor und nach der Anwendung der KS-Regeln fehlerfrei); Variablentyp: ordinal.
- Statistische Auswertung: Häufigkeitstabelle auf Basis der Fehlerannotation

5.3.1.3 Dritter Analysefaktor: Aufteilung der Annotationsgruppen

- In der Studie werden die Ergebnisse der Fehlerannotation in vier Annotationsgruppen unterteilt:

  (1) RR: MÜ ist vor und nach der Anwendung der KS-Regel fehlerfrei; (2) FF: MÜ beinhaltet vor und nach der Anwendung der KS-Regel Fehler; (3) RF: MÜ ist nur vor der Anwendung der KS-Regel fehlerfrei; (4) FR: MÜ ist nur nach der Anwendung der KS-Regel fehlerfrei.
- Fragestellung: Wie hoch ist der Prozentsatz jeder Annotationsgruppe?
- Statistische Auswertung: Häufigkeiten mit Bootstrapping[2] auf Basis der Fehlerannotation; Abbildungen: Kreisdiagramm

5.3.1.4 Vierter Analysefaktor: Vergleich der Fehlertypen vor vs. nach der Anwendung der KS-Regeln

- Es werden die 13 analysierten Fehlertypen vor vs. nach der Anwendung der KS-Regeln verglichen.
- Fragestellung: Kommen bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regeln vor?

[2] Bootstrapping ist ein statistisches Verfahren zur Schätzung der Stichprobenverteilung eines Schätzers durch erneute Stichprobenerstellung mit Ersatz aus der ursprünglichen Stichprobe. Es wird als ein nützliches Verfahren zum Testen der Modellstabilität betrachtet. (IBM o.D.)

Davon wird abgeleitet, (1) ob bestimmte Fehlertypen, die vor der Anwendung der KS-Regeln existierten, nach der Anwendung der KS-Regeln eliminiert bzw. reduziert wurden; (2) ob bestimmte Fehlertypen erst nach der Anwendung der KS-Regeln auftraten bzw. deutlich stiegen (im Vergleich zu vor der Anwendung der KS-Regeln).

- Variablen: Fehler existiert ja/nein; Fehlertyp: dichotom. Summe der Fehler sowie Mittelwert der Fehleranzahl innerhalb der KS-Stelle; Variablentyp: ordinal
- Statistische Auswertung: Kreuztabellen auf Basis der Fehlerannotation; Abbildungen: Balken
- Hypothesen:

  H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

  H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: McNemar; Begründung der Testauswahl: Mithilfe des McNemar-Tests können zwei verbundene dichotome Parameter verglichen werden, somit kann eine mögliche signifikante Veränderung bei einem Fehlertyp vor vs. nach der Anwendung der KS-Regeln identifiziert werden.

5.3.1.5 Fünfter Analysefaktor: Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln

- Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?
- Variablen: Mittelwert der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch
- Statistische Auswertung: Deskriptive Statistiken auf Basis der Humanevaluation; Abbildungen: Fehlerbalken
- Hypothesen:

  H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

5.3.1.6 Sechster Analysefaktor: Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln auf Annotationsgruppenebene

- Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

  Davon wird abgeleitet, (1) ob bei der Gruppe RR die Stil- bzw. Inhaltsqualität vor bzw. nach der Anwendung der KS-Regeln höher ist, obwohl die MÜ in beiden Fällen fehlerfrei ausfällt; (2) ob bei der Gruppe FF die Stil- bzw. Inhaltsqualität vor bzw. nach der Anwendung der KS-Regeln höher ist, obwohl die MÜ in beiden Fällen Fehler beinhaltet; (3) ob bei der Gruppe RF die Stil- bzw. Inhaltsqualität nach der Anwendung der KS-Regeln stieg, obwohl die MÜ nach der Anwendung der KS-Regeln Fehler beinhaltet und davor fehlerfrei war; (4) ob bei der Gruppe FR die Stil- bzw. Inhaltsqualität nach der Anwendung der KS-Regeln sank, obwohl die MÜ nach der Anwendung der KS-Regeln fehlerfrei war und davor Fehler beinhaltete.

- Variablen: Mittelwert der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala in jeder Annotationsgruppe; Variablentyp: metrisch

- Statistische Auswertung: Deskriptive Statistiken auf Basis der Humanevaluation unter Aufteilung der Daten nach den Annotationsgruppen; Abbildungen: Fehlerbalken

- Hypothesen:

  H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

  H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

5.3.1.7 Siebter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

- Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS *minus* Qualität vor KS)?
- Variablen: Differenz der Mittelwerte der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch. Differenz der Fehleranzahl der einzelnen Fehlertypen; Variablentyp: ordinal.
- Statistische Auswertung: Spearman-Korrelationstest
- Hypothesen:

  H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

  H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

- Signifikanztest: Spearman-Korrelationstest; Begründung der Testauswahl: Eine der Variablen ist ordinal. Zudem setzt Spearman keine Anforderung an die Verteilung und die Linearität voraus.

5.3.1.8 Achter Analysefaktor: Vergleich der AEM-Scores vor vs. nach der Anwendung der KS-Regeln

- Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?
- Variablen: Mittelwert der AEM-Scores[3] sowie Differenzen der AEM-Scores (AEM-Score nach KS *minus* AEM-Score vor KS); Variablentyp: metrisch

[3] Bei jeder MÜ wurde der Mittelwert der AEM-Scores auf Basis von zwei Referenzübersetzungen ermittelt; für eine genaue Beschreibung des Verfahrens siehe §4.5.6.4.

- Statistische Auswertung: Deskriptive Statistiken auf Basis der automatischen Evaluation; Abbildungen: Fehlerbalken für die Differenzen der AEM-Scores[4]
- Hypothesen:

  H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

  H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.
- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

5.3.1.9 Neunter Analysefaktor: Korrelation zwischen den AEM-Scores-Differenzen und der Qualitätsdifferenz

- Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores in TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS *minus* Mittelwert der AEM-Scores vor KS) und der Differenz der allgemeinen Qualität[5] (Qualität nach KS *minus* Qualität vor KS)?
- Variablen: Differenz der Mittelwerte der AEM-Scores sowie Differenz der Mittelwerte der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch
- Statistische Auswertung: Spearman-Korrelationstest
- Hypothesen:

  H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

[4] Bei der Auswertung werden nur die Differenzen der AEM-Scores (und nicht die Mittelwerte der AEM-Scores) verwendet. Der Grund dafür ist, dass die Bewerter die komplette MÜ editiert haben. Ihre Edits können daher Stellen außerhalb der KS-Stelle umfassen. Da aber die MÜ vor und nach KS außerhalb der KS-Stelle vereinheitlicht wurden, wird hier davon ausgegangen, dass wir durch die Verwendung der Differenz (AEM-Score nach KS − AEM-Score vor KS) nur die Edits innerhalb der KS-Stelle betrachten; für eine detaillierte Erläuterung siehe §4.5.6.4.

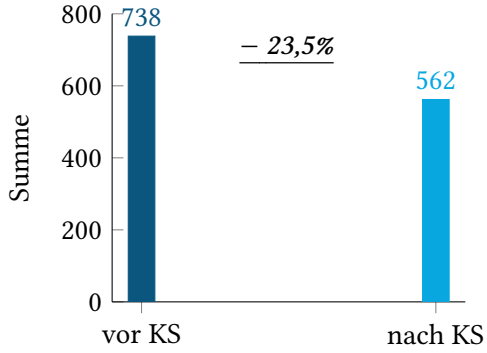
[5] Die allgemeine Qualität ist der Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Untersuchung dieser Korrelation keine Unterscheidung zwischen der Stil- und Inhaltsqualität notwendig ist.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

- Signifikanztest: Spearman-Korrelationstest; Begründung der Testauswahl: Nicht alle Variablen sind normalverteilt. Spearman setzt keine Anforderung an die Verteilung und die Linearität voraus.

5.3.2 Vergleich der Fehleranzahl vor vs. nach der Anwendung aller KS-Regeln

Die Fehleranzahl sank system- und regel-übergreifend signifikant um knapp 23,5 % (Abbildung 5.2) von 738 Fehlern vor der Anwendung der KS-Regeln (M = ,68 / SD = ,898 / N = 1080) auf 562 Fehler nach der Anwendung der KS-Regeln (M = ,52 / SD = ,825 / N = 1080) (Abbildung 5.3).[6]



Signifikante Diff.

Abbildung 5.2: Summe der Fehleranzahl vor vs. nach KS auf Sprachenpaarebene

Der Mittelwert der Differenz (nach KS minus vor KS) in der Fehleranzahl pro Satz lag bei – ,16 (SD = ,961) mit einem 95%-Konfidenzintervall[7] zwischen einem

[6] Abbildung 5.3 ist wie folgt zu lesen: Die Punkte zeigen, wie hoch die durchschnittliche Fehleranzahl ausfällt (z. B. ,68 vor KS). Die Fehlerbalken zeigen das realisierte 95 %-Konfidenzintervall (CI) für die durchschnittliche Fehleranzahl (in dem Fall beläuft sich das CI vor KS auf ,63; ,74). Demnach würde die Fehleranzahl bei der Durchführung einer weiteren vergleichbaren Untersuchung mit einer Wahrscheinlichkeit von 95 % zwischen ,63 und ,74 liegen (vgl. Eckstein 2008: 81).

[7] Konfidenzintervalle bezeichnen „Intervalle mit einer Ober- und einer Untergrenze. Sie geben die Sicherheit der Schätzung einer gesuchten Kenngröße, z. B. des Mittelwerts, an“ (Keller 2015).

Minimum von − ,22 (SD = ,901) und einem Maximum von −0, 10 (SD = 1,026) (Bootstrapping mit 1000 Stichproben).[8] Die Differenz (nach KS – vor KS) in der Fehleranzahl erwies sich als hochsignifikant ($z(N = 1080) = -5,589, p < 0,001$).

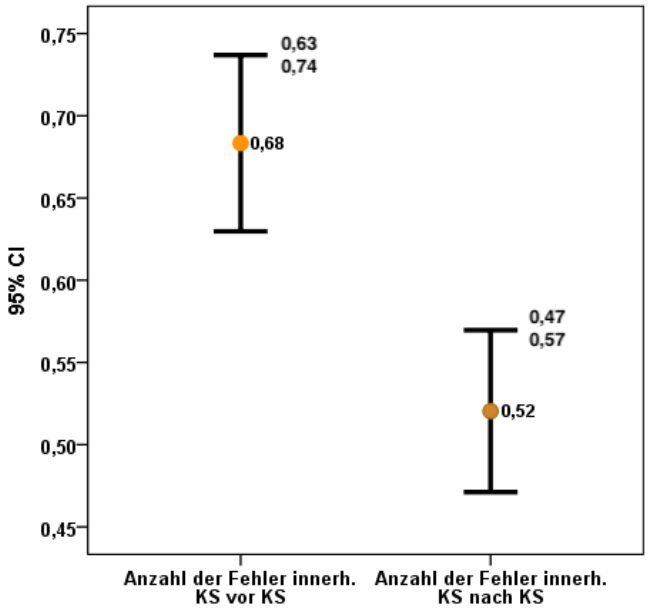


Abbildung 5.3: Mittelwert der Fehleranzahl pro Satz vor vs. nach KS auf Sprachenpaarebene

5.3.3 Vergleich der Fehleranzahl vor vs. nach KS außerhalb der KS-Stelle bei der Gruppe RR

Ein Szenario, das nicht außer Acht gelassen durfte, war die Untersuchung, ob die KS-Stelle vor und nach der Anwendung der KS-Regeln korrekt übersetzt wurde *und gleichzeitig* die Fehleranzahl *außerhalb* der KS-Stelle nach der Anwendung der KS-Regeln stieg. In der folgenden Häufigkeitstabelle (Tabelle 5.7) werden die Ergebnisse präsentiert (N = 490).

[8] Bootstrapping ist eine Resampling-Methode, bei der aus der Stichprobe viele Stichproben (in der vorliegenden Studie 1000 Stichproben) entnommen werden. Die Datenanalyse wird dann jeweils auf Grundlage der vielen Stichproben durchgeführt und die Ergebnisse werden zu einem Endergebnis zusammengefasst und mit dem Konfidenzintervall (CI) (in dieser Studie mit CI 95 %) angegeben. Ziel ist es, ein verlässliches Endergebnis zu erhalten, auch wenn die Daten nicht normalverteilt sind oder Ausreißer beinhalten (Keller 2019). Mehr zu Bootstrapping in Chernick (2008).

Tabelle 5.7: Häufigkeit der Differenz der Fehleranzahl außerhalb der KS-Stelle bei einer korrekten Übersetzung der KS-Stelle auf Sprachenpaarebene

Differenz der Fehleranzahl nach KS *minus* vor KS außerhalb KS	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
+ 3	3	,6	,6	,6
+ 2	2	,4	,4	1,0
+ 1	27	5,5	5,5	6,5
0	413	84,3	84,3	90,8
− 1	30	6,1	6,1	96,9
− 2	10	2,0	2,0	99,0
− 3	4	,8	,8	99,8
− 5	1	,2	,2	100,0
Gesamt	490	100,0	100,0	

Tabelle 5.7 listet die Differenzen in der Fehleranzahl aller Übersetzungen auf, die innerhalb der KS-Stelle vor und nach der Anwendung der KS-Regeln fehlerfrei waren, aber eine Veränderung in der Fehleranzahl außerhalb der KS-Stelle hatten. Nach den demonstrierten Häufigkeiten der Differenzen gab es keinen auffälligen Anstieg oder Abstieg bei der Fehleranzahl.

Ferner wurden die Daten getrennt auf Regelebene und auf Systemebene untersucht, um herauszufinden, ob ein auffälliger Anstieg bei einer bestimmten Regel bzw. einem bestimmten System vorkam. Auf Regelebene gab es keine hohen Häufigkeiten. Auf Systemebene stieg nur bei Google Translate und Bing die Fehleranzahl um + 1 Fehler: bei Google Translate 10 Mal innerhalb von 183 Fällen und bei Bing 12 Mal innerhalb von 81 Fällen. Es handelt sich hierbei um vereinzelte Anstiege, die bei verschiedenen KS-Regeln vorkamen (mehr dazu unter §5.5.3).

5.3.4 Aufteilung der Annotationsgruppen

Aus der Untersuchung der Aufteilung der Annotationsgruppen ergab sich Folgendes (Abbildung 5.4):

Knapp 45 % der Übersetzungen waren sowohl vor als auch nach der Anwendung der KS-Regeln korrekt (Gruppe RR). Außerdem waren fast 26 % der Übersetzungen vor und nach der Anwendung der KS-Regeln (Gruppe FF) falsch. Im

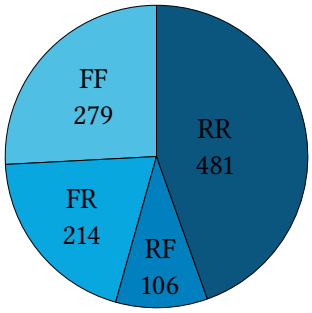


Abbildung 5.4: Aufteilung der Annotationsgruppen auf Sprachenpaarebene

nächsten Abschnitt werden die persistenten Fehlertypen ins Visier genommen. Wesentlich im Ergebnis ist, dass ungefähr 20 % der vor KS falschen Übersetzungen nach KS korrigiert wurden (Gruppe FR). Gleichzeitig gab es Fälle (ca. 10 %), die nur nach KS falsch übersetzt wurden (Gruppe RF) (Abbildung 5.4).

Die Ermittlung der Konfidenzintervalle (CI 95%) der Aufteilung der Annotationsgruppen mithilfe eines Bootstrapping mit 1000 Stichproben ergab die folgenden Unter- und Oberwerte (N = 1080):

Tabelle 5.8: Konfidenzintervalle (CI 95%) der Aufteilung der Annotationsgruppen auf Sprachenpaarebene

Annotationsgruppe	Häufigkeit	Prozente	Bootstrapping		
				95%-Konfidenzintervall	
			Verzerrung	Unterer Wert	Oberer Wert
FF	279	25,8	,0	23,1	28,5
FR	214	19,8	,0	17,4	22,3
RF	106	9,8	,0	8,1	11,6
RR	481	44,5	,0	41,5	47,5
Gesamt	1080	100,0	,0	100,0	100,0

Durch das 95%-Konfidenzintervall wird belegt, dass bei der Durchführung einer weiteren vergleichbaren Untersuchung die Aufteilung der Annotationsgruppen mit einer Wahrscheinlichkeit von 95 % zwischen den aufgeführten Unter- und Oberwerten liegen würde.

5.3.5 Vergleich der Fehlertypen vor vs. nach der Anwendung aller KS-Regeln

Die Veränderungen in der Fehleranzahl bei den verschiedenen Fehlertypen nach der Anwendung der KS-Regeln variieren zwischen gestiegen, gesunken und fast gleichgeblieben (Abbildung 5.5). Eine positive Wirkung hatten die KS-Regeln am stärksten auf den semantischen Fehlertyp SM.13 „Kollokation“ (− 68 %) gefolgt von dem lexikalischen Fehlertyp LX.3 „Wort ausgelassen“ (− 61 %). Als nächstes sank der grammatische Fehler GR.8 „Falsches Verb“ (− 56 %) gefolgt vom orthografischen OR.2 „Großschreibfehler“ (− 52 %). Weitere Rückgänge der Fehleranzahl waren bei dem GR.10 „Wortstellungsfehler“ (− 38 %) und dem GR.9 „Kongruenzfehler“ (− 21 %) zu beobachten. Zudem nahmen, wie die Abbildung 5.5 zeigt, einige Fehlertypen nach der Anwendung der KS-Regeln zu. Dies ist insbesondere beim LX.6 „Konsistenzfehler“ (+ 700 %) und dem orthografischen OR.1 „Zeichensetzungsfehler“ (+ 186 %) der Fall.

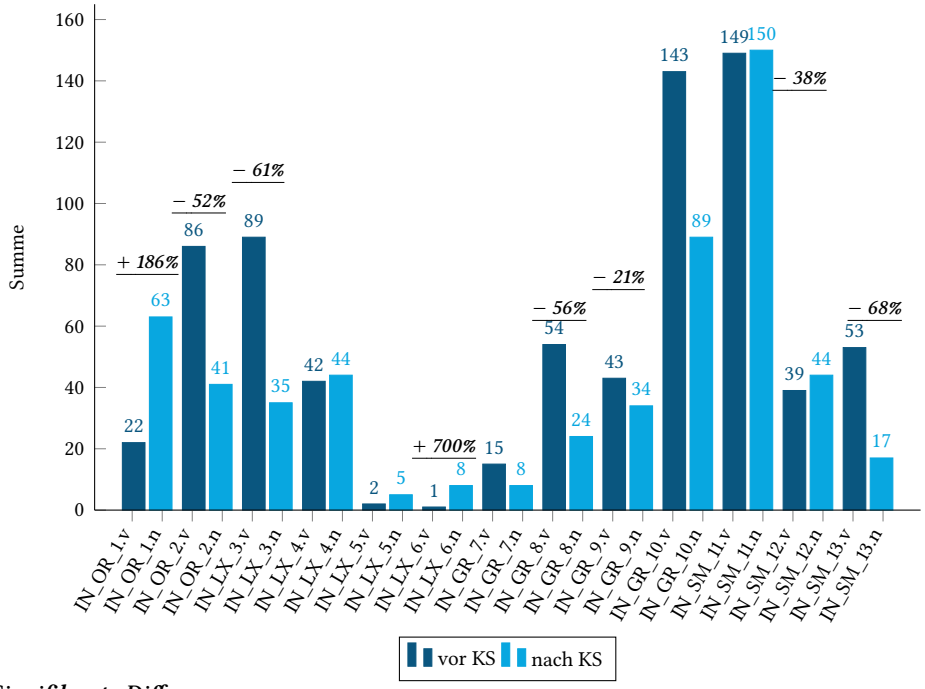
Nur bei den folgenden Fehlertypen erwies sich die Differenz in der Fehleranzahl als signifikant (Tabelle 5.9).

Die acht signifikant veränderten Fehlertypen entstammen allen linguistischen Bereichen (Orthografie, Lexik, Grammatik sowie Semantik). Mit Ausnahme von Fehlertyp LX.6 „Lexik – Konsistenzfehler“ waren die Differenzen bei allen Fehlertypen hochsignifikant. Wie die Analyse auf Regelebene zeigt, kam die signifikante Veränderung bei Fehlertyp LX.6 fast nur bei Regel „Eindeutige pronominale Bezüge verwenden“ vor (siehe §5.4.5.4).[9] Dies begründet die kleine Fehleranzahl und entsprechend das niedrigere Signifikanzniveau. Innerhalb der acht signifikant veränderten Fehlertypen stieg die Fehleranzahl nur bei zwei Fehlertypen (OR.1 „Zeichensetzung“ bei der Regel „Partizipialkonstruktionen vermeiden“ (siehe §5.4.6.4) und LX.6 „Konsistenz“ bei der Regel „Eindeutige pronominale Bezüge verwenden“ (siehe §5.4.5.4) und sank bei den übrigen sechs Fehlertypen. Mögliche Interpretationen hierfür sind regelspezifisch und daher auf Regelebene präsentiert.

5.3.6 Vergleich der MÜ-Qualität vor vs. nach der Anwendung aller KS-Regeln

Ein Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln zeigt, dass sowohl die Stil- als auch die Inhaltsqualität stiegen (Abbildung 5.6):

[9] Außerdem kam Fehlertyp LX.6 bei einer einzigen Übersetzung bei der Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ vor.



Signifikante Diff.
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Wort zusätzlich falsch eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.5: Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS auf Sprachenpaarebene

Tabelle 5.9: Fehlertypen mit signifikanter Veränderung nach der KS-Anwendung auf Sprachenpaarebene

	N	Mittelwert	Standard-abweichung	Signifikanz (McNemar-Test)
OR.1: Orthografie – Zeichensetzung	1080	vor KS = ,02 nach KS = ,06	vor KS = ,165 nach KS = ,278	p < ,001
OR.2: Orthografie – Großschreibung	1080	vor KS = ,08 nach KS = ,04	vor KS = ,271 nach KS = ,191	p < ,001
LX.3: Lexik – Wort ausgelassen	1080	vor KS = ,08 nach KS = ,03	vor KS = ,282 nach KS = ,187	p < ,001
LX.6: Lexik – Konsistenzfehler	1080	vor KS = ,00 nach KS = ,01	vor KS = ,030 nach KS = ,086	p = ,039
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)	1080	vor KS = ,05 nach KS = ,02	vor KS = ,218 nach KS = ,147	p < ,001
GR.9: Grammatik – Kongruenzfehler	1080	vor KS = ,04 nach KS = ,03	vor KS = ,196 nach KS = ,213	p = ,002
GR.10: Grammatik – Falsche Wortstellung	1080	vor KS = ,13 nach KS = ,08	vor KS = ,365 nach KS = ,282	p < ,001
SM.13: Semantik – Kollokationsfehler	1080	vor KS = ,05 nach KS = ,02	vor KS = ,225 nach KS = ,125	p < ,001

Kursiv = gestiegene Fehlertypen nach KS

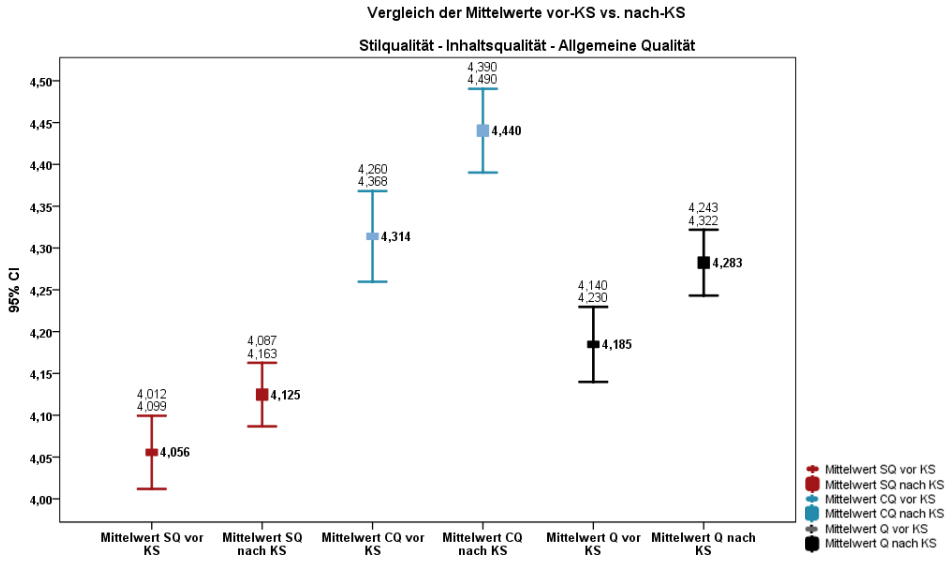


Abbildung 5.6: Mittelwerte der Qualität vor vs. nach KS auf Sprachenpaarebene

Die Stilqualität stieg um 1,7 % (Mv = 4,056 / SDv = ,620 / Mn = 4,125 / SDn = ,539 / N = 775). Die Inhaltsqualität stieg um 2,9 % (Mv = 4,314 / SDv = ,769 / Mn = 4,440 / SDn = ,710 / N = 775). Der Mittelwert der Differenz (nach KS minus vor KS) in den vergebenen Qualitätspunkten pro Satz lag für die Stilqualität bei ,069 (SD = ,637) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,024 und einem Maximum von ,114 und für die Inhaltsqualität bei ,126 (SD = ,830) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,068 und einem Maximum von ,185 (Bootstrapping mit 1000 Stichproben) (Abbildung 5.7). Das 95%-Konfidenzintervall besagt, dass bei der Durchführung einer weiteren vergleichbaren Untersuchung der Mittelwert der Differenz mit einer Wahrscheinlichkeit von 95 % zwischen den aufgeführten Unter- und Oberwerten liegen würde (vgl. Eckstein 2008: 81).

Die Differenzen (nach KS *minus* vor KS) in der Stil- und Inhaltsqualität erwiesen sich als signifikant (z (N = 775) = – 2,062 / p = ,039) bzw. (z (N = 775) = – 4,566 / p < ,001). Dieser allgemeine positive Einfluss der KS auf den MÜ-Output wurde in anderen empirischen und theoretischen Studien bestätigt (vgl. Nyberg & Mitamura 1996; Bernth 1999; Bernth & Gdaniec 2001: 208; Drugan 2013: 98; Drewer & Ziegler 2014: 196; Wittkowsky 2017: 92). Dennoch ist es beachtenswert, dass sich die Inhaltsqualität im Vergleich zu der Stilqualität nach KS wesentlich verbesserte. Das wirft die Frage auf, in genau welchen Fällen die Inhaltsqualität mehr als

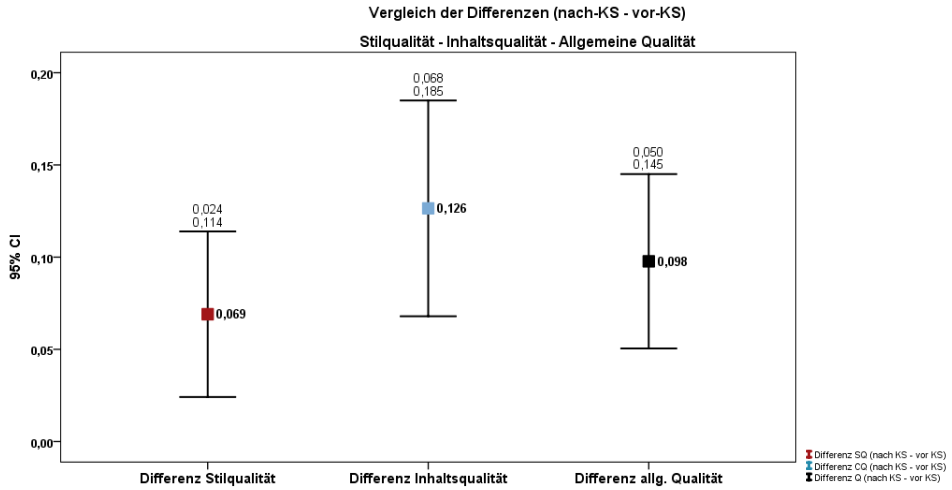


Abbildung 5.7: Mittelwert der Qualitätsdifferenzen auf Sprachenpaarebene

die Stilqualität stieg. Dies lässt sich auf Regelebene beantworten (Abbildung 6.3): Bei den fünf Regeln „Konditionalsätze als ‚Wenn'-Sätze formulieren", „Eindeutige pronominale Bezüge verwenden", „Partizipialkonstruktionen vermeiden", „Überflüssige Präfixe vermeiden" und „Keine Wortteile weglassen" stieg die Inhaltsqualität nach der Anwendung der KS mehr als die Stilqualität, während bei den drei Regeln „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden", „Funktionsverbgefüge vermeiden", „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden" der Anstieg der Stilqualität höher war (detailliert erläutert auf Regelebene, siehe §5.4). Bei der neunten Regel „Passiv vermeiden" waren die Differenzen bei der Stil- und Inhaltsqualität vergleichbar.

5.3.7 Vergleich der MÜ-Qualität vor vs. nach der Anwendung aller KS-Regeln auf Annotationsgruppenebene

Eine genaue Betrachtung der Annotationsgruppen zeigt Folgendes (Abbildung 5.8): Erwartungsgemäß sanken die Stil- und Inhaltsqualität bei der Gruppe RF (Übersetzung vor KS richtig und nachher falsch) signifikant (N = 83 / p < ,001) und stiegen bei der Gruppe FR (Übersetzung vor KS falsch und nachher richtig) signifikant (N = 159 / p < ,001).

Bei der Gruppe FF (falsch vor und nach KS) gab es eine minimale Steigerung, die sich nicht als signifikant erweisen konnte. Hingegen liefert die Gruppe RR (richtig vor und nach KS) ein wesentliches Ergebnis: Die Stilqualität nach KS

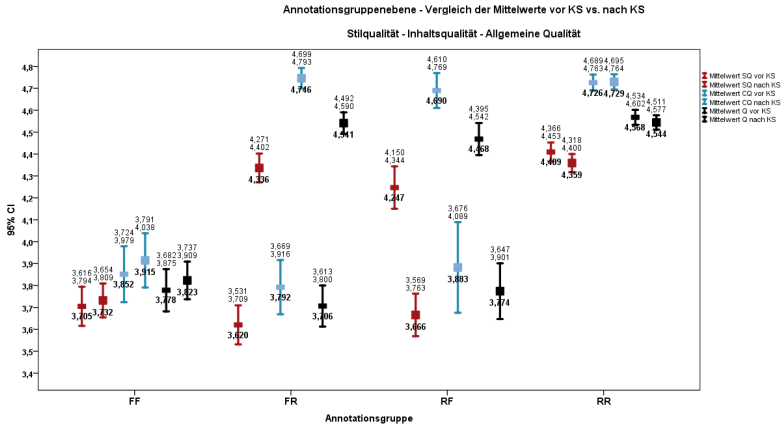


Abbildung 5.8: Mittelwerte der Qualität vor vs. nach KS bei den einzelnen Annotationsgruppen auf Sprachenpaarebene

fällt signifikant geringer als vor KS (N = 341 / p = ,022) aus. Auf der anderen Seite blieb die Inhaltsqualität fast unverändert. In anderen Worten vergleicht man zwei fehlerfreie MÜ vor und nach der Anwendung der KS, wäre die Stilqualität vor KS besser als nach KS, während die Inhaltsqualität vergleichbar bliebe. Aus stilistischer Sicht – wie die Ergebnisse auf Regelebene (siehe §6.3) zeigen – fanden die Bewerter die Verwendung einiger Regeln unnatürlich bzw. unidiomatisch.

Zudem wurde nur ein extrem niedriges Konfidenzintervall (Abbildung 5.9) bei der Gruppe RR registriert. Dies deutet darauf hin, dass bei der Durchführung einer weiteren Untersuchung vom gleichen Umfang und aus der gleichen Grundgesamtheit die Qualitätswerte mit einer Wahrscheinlichkeit von 95 % sehr ähnlich ausfallen würden.

5.3.8 Korrelation zwischen der Differenz der Fehlertypen und den Qualitätsdifferenzen

Die Spearman-Korrelation erwies einen signifikanten negativen mittleren Zusammenhang zwischen der Differenz im LX.3 „Lexik – Wort ausgelassen“ und der Differenz in der Inhaltsqualität sowie zwischen der Differenz im GR.10 „Grammatik – Falsche Wortstellung“ und der Differenz in der Stil- und Inhaltsqualität. Erwartungsgemäß profitierten beide Qualitätsattribute von der 38 %-igen Minderung der Wortstellungsfehler (GR.10), siehe Abbildung 5.5. Ebenfalls beeinflusste das Auslassen von Wörtern (LX.3) vor der Anwendung der KS-Regeln die Inhaltsqualität negativ. Nach der Anwendung der KS-Regeln wurde dieser Fehlertyp in

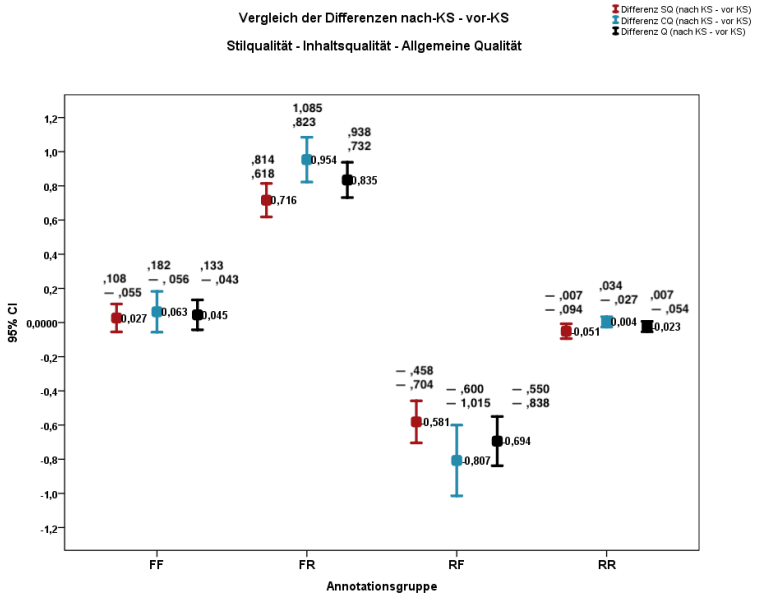


Abbildung 5.9: Qualitätsdifferenzen vor- vs. nach KS auf Annotationsgruppenebene

61 % der Fälle korrigiert (siehe Abbildung 5.5), was mit einem Anstieg der Inhaltsqualität verbunden war (Tabelle 5.10).

Weitere mittlere oder starke Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht belegt werden (Tabelle 5.10).

5.3.9 Vergleich der AEM-Scores vor vs. nach der Anwendung aller KS-Regeln

Durchschnittlich verbesserten sich sowohl der TERbase-Score als auch der hLEPOR-Score[10] minimal nach der Anwendung der KS-Regeln im Vergleich zu vorher:[11] Die Differenz (nach KS *minus* vor KS) lag beim TERbase-Score bei ,0011 und beim hLEPOR-Score bei ,0087 (Abbildung 5.10).

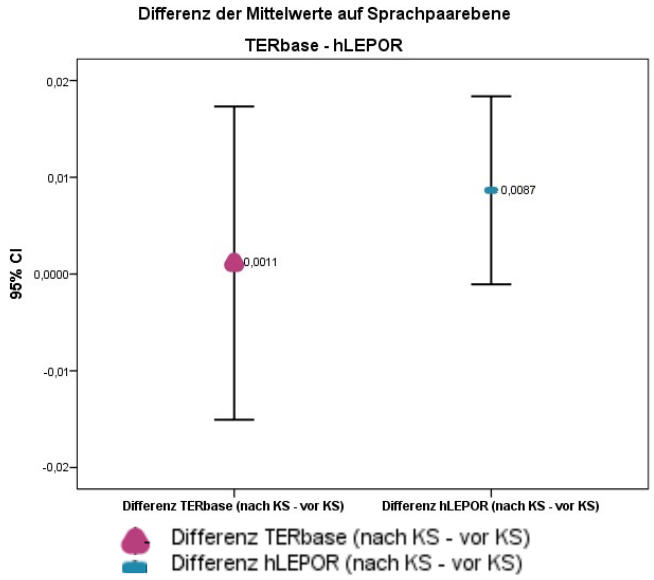
[10] Näheres zu den AEMs TERbase und hLEPOR unter §4.5.6.3 „Auswahl der automatischen Evaluationsmetriken".

[11] Eine genaue Erklärung für die Berechnung des Mittelwerts der Differenz des TERbase-Scores und hLEPOR-Scores sowie seine Bedeutung ist unter §4.5.6.4 „Basis des Vergleichs vor-KS vs. nach-KS zur Ermittlung des KS-Einflusses" aufgeführt.

Tabelle 5.10: Korrelationen zwischen den Fehlertypen und der Qualität auf Sprachenpaarebene

	N	Signifikanz (p)			Korrelationskoeffizient (ρ)		
		SQ	CQ	Q	SQ	CQ	Q
Differenz der Anzahl der **OR.1**	775	< ,001	,019	< ,001	− ,212	− ,084	− ,170
Differenz der Anzahl der **OR.2**	775	< ,001	,026	< ,001	− ,170	−,080	− ,130
Differenz der Anzahl der **LX.3**	775	< ,001	< ,001	< ,001	− ,204	− ,366	− ,321
Differenz der Anzahl der **LX.4**	775	< ,001	< ,001	< ,001	− ,235	− ,224	− ,240
Differenz der Anzahl der **LX.5**	775	,059	,003	,007	− ,068	− ,107	− ,097
Differenz der Anzahl der **LX.6**	775	,029	,011	,012	− ,078	− ,091	− ,090
Differenz der Anzahl der **GR.7**	775	,513	,004	,083	− ,024	− ,102	− ,062
Differenz der Anzahl der **GR.8**	775	< ,001	< ,001	< ,001	− ,269	− ,239	− ,264
Differenz der Anzahl der **GR.9**	775	,070	,007	,016	− ,065	− ,097	− ,087
Differenz der Anzahl der **GR.10**	775	< ,001	< ,001	< ,001	− ,319	− ,307	− ,337
Differenz der Anzahl der **SM.11**	775	< ,001	< ,001	< ,001	− ,206	− ,230	− ,246
Differenz der Anzahl der **SM.12**	775	,006	< ,001	< ,001	− ,099	− ,226	− ,203
Differenz der Anzahl der **SM.13**	775	,002	,239	,022	− ,113	− ,042	− ,082

schwache Korrelation (ρ >= 0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5) nicht signifikant (p > 0,05)



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.10: Mittelwert der Differenz des TERbase-Scores und hLEPOR-Scores auf Sprachenpaarebene

5.3.10 Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

Die Spearman-Korrelation erwies einen hochsignifikanten positiven starken Zusammenhang zwischen den Differenzen der TERbase-Scores und hLEPOR-Scores und der Differenz der Qualität im Allgemeinen (Tabelle 5.11).[12] Dies besagt: Verbessert sich der AEM-Score, verbessert sich ebenfalls die Qualität.

Somit bestätigen sich die Ergebnisse der Humanevaluation und die der automatischen Evaluation gegenseitig, denn eine starke positive Korrelation besagt, dass ein positiver Effekt der KS-Anwendung sich sowohl durch einen erhöhten Score in der Humanevaluation als auch einen verbesserten AEM-Score in der automatischen Evaluation zeigte. Umgekehrt gingen auch bei einem negativen Effekt der KS-Anwendung die Scores beider Evaluationen zurück.

[12] Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Ermittlung dieser Korrelation keine Unterscheidung zwischen den beiden Qualitätsattributen notwendig ist.

Tabelle 5.11: Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelations-koeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	775	< ,001	,520	starker Zusammenhang
Korrelation zw. Differenz der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	775	< ,001	,519	starker Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

5.3.11 Analyse auf Sprachenpaarebene – Validierung der Hypothesen

Um die in diesem Unterkapitel dargestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der obigen Ergebnisse in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (−) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[13]

Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.

[13] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.

Resultat Die Fehleranzahl sank nach der Anwendung der KS-Regeln signifikant. Somit wurde H0 abgelehnt und H1 bestätigt.

Vergleich der Fehlertypen vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Kommen bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regeln vor?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

Resultat H0 wurde abgelehnt und somit H1 bestätigt, und zwar nur für die hier aufgeführten Fehlertypen: + OR.1, − OR.2, − LX.3, + LX.6, − GR.8, − GR.9, − GR.10, − SM.13

Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

Resultat Die Stil- und Inhaltsqualität stiegen nach der Anwendung der KS-Regeln signifikant. Somit wurde H0 abgelehnt und H1 bestätigt.

Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

Resultat In den folgenden Fällen wurde H0 abgelehnt und somit H1 bestätigt: bei dem Qualitätsrückgang in der Gruppe RF, bei der Qualitätssteigerung in FR und bei der Abnahme der Stilqualität in RR.

FF		RF		FR		RR	
SQ	CQ	SQ	CQ	SQ	CQ	SQ	CQ
+	+	(−)	(−)	+	+	(−)	+

Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS *minus* Qualität vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat H0 wurde abgelehnt und H1 bestätigt, und zwar für den Zusammenhang zwischen den Qualitätswerten und zwei Fehlertypen wie folgt:
neg GR.10 <> SQ, *neg* LX.3 <> CQ, *neg* GR.10 <> CQ

Vergleich der AEM-Scores vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

Resultat Die AEM-Scores von TERbase bzw. hLEPOR verbesserten sich nur minimal. Somit wurde H0 nicht abgelehnt und H1 konnte nicht bestätigt werden.

Korrelation zwischen den AEM-Scores-Differenzen und der Qualitätsdifferenz

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores in TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS *minus* Mittelwert der AEM-Scores vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS *minus* Qualität vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat H0 wurde abgelehnt und H1 bestätigt, und zwar für den Zusammenhang zwischen der allgemeinen Qualität und den AEM-Scores wie folgt: *pos* TERbase <<>> Q, *pos* hLEPOR <<>> Q

5.3.12 Übersicht der Ergebnisse auf Sprachenpaarebene

Tabelle 5.12 bietet eine Übersicht über die Ergebnisse auf Sprachenpaarebene, d. h. für den kompletten Datensatz.

Regel- und systemübergreifend zeigt die Anwendung der neun KS-Regeln einen positiven Einfluss auf den MÜ-Output. Die Fehleranzahl sank und die Stil- und Inhaltsqualität stiegen signifikant. Die AEM-Scores verbesserten sich leicht. Die signifikant starke positive Spearman-Korrelation zwischen beiden AEMs und

Tabelle 5.12: Übersicht der Ergebnisse auf Sprachenpaarebene

Fehler-anzahl	Fehlertypen	Qualität		Fehlertypen <> Qualität		AEM-Scores		AEM-Scores <> allg. Qualität	
				Stilqualität	Inhalts-qualität	TERbase	hLEPOR	TERbase	hLEPOR
(–)	+ OR.1 – OR.2 – LX.3 + LX.6 – GR.8 – GR.9 – GR.10 – SM.13	+ SQ	+ CQ	*neg* GR.10 <> SQ	*neg* LX.3 <> CQ *neg* GR.10 <> CQ	+	+	*pos* TERbase <<>> Q	*pos* hLEPOR <<>> Q

SQ: Stilqualität CQ: Inhaltsqualität Q: allg. Qualität Signifikant ($p < 0{,}5$) Blank: nicht signifikant
<> mittlere Korrelation ($\rho >= 0{,}3$) <<>> starke Korrelation ($\rho >= 0{,}5$) *neg*: negative Korrelation *pos*: positive Korrelation

der allgemeinen Qualität lässt beide Analysen (Humanevaluation und automatische Evaluation) den positiven Effekt der KS-Regeln bestätigen. Die Fehleranzahl mehrerer Fehlertypen sank signifikant nach der KS-Anwendung. Diese sind: OR.2 „Großschreibung“, LX.3 „Wort ausgelassen“, GR.8 „Falsches Verb (Zeitform, Komposition, Person)“, GR.9 „Kongruenzfehler“, GR.10 „Falsche Wortstellung“ und SM.13 „Kollokationsfehler“. Gleichzeitig stieg die Fehleranzahl zweier Fehlertypen OR.1 „Zeichensetzung“ (überwiegend durch die Regel „Partizipialkonstruktionen vermeiden“, siehe §5.4.6.4) und LX.6 „Konsistenzfehler“ (überwiegend durch die Regel „Eindeutige pronominale Bezüge verwenden“, siehe §5.4.5.4). Den Ergebnissen des Spearman-Tests zufolge besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp GR.10 „Falsche Wortstellung“ und der SQ und CQ sowie zwischen LX.3 „Wort ausgelassen“ und der CQ. Somit war der Rückgang dieser Fehlertypen mit einem Anstieg der genannten Qualitätswerte verbunden.

5.4 Analyse auf Regelebene sowie auf Regel- und MÜ-Systemebene

In diesem Kapitel werden die Ergebnisse auf Regelebene sowie auf Regel- und MÜ-Systemebene dargestellt. Auf Regelebene wird jede Regel einzeln anhand einer Reihe von Faktoren analysiert, die den Vergleich der Szenarien „vor der Anwendung der KS-Regel“ vs. „nach der Anwendung der KS-Regel“ ermöglichen. Anschließend werden unter jeder Regel die fünf MÜ-Systeme verglichen (Regel- und MÜ-Systemebene). Für den Vergleich werden folgende Analysefaktoren herangezogen: Fehleranzahl, Aufteilung der Annotationsgruppen, Fehlertypen, Stil- und Inhaltsqualität, Korrelation zwischen den Fehlertypen und der Stil- und Inhaltsqualität, Stil- und Inhaltsqualität auf Annotationsgruppenebene, AEM-Scores und Korrelation zwischen den Differenzen der AEM-Scores und der Qualitätsdifferenz.[14] Bevor die Ergebnisse demonstriert werden, gibt uns der erste Abschnitt eine Übersicht über die Analysefaktoren, die zugrundeliegenden Fragestellungen und Hypothesen sowie die statistische Auswertung. Danach folgt die quantitative und qualitative Analyse der einzelnen Regeln dargestellt in alphabetischer Reihenfolge. Bezüge auf vorherige Studien werden nicht in diesem Unterkapitel, sondern im Rahmen der Diskussion in §6 vorgenommen.

[14] Zur Berechnung aller Differenzen in der Studie wird durchgehend das Szenario „nach KS“ minus „vor KS“ subtrahiert, somit ist die Qualitätsdifferenz = Qualität nach KS − Qualität vor KS; AEM-Score-Differenz = AEM-Score nach KS − AEM-Score vor KS usw.

5.4.1 Analysefaktoren

Für den Vergleich der Übersetzungen vor vs. nach der Anwendung jeder KS-Regel wurden die Übersetzungen *außerhalb* der KS-Stelle vereinheitlicht, somit war nur die KS-Stelle in den beiden Szenarien unterschiedlich (für die genaue Vorgehensweise siehe §4.5.3.1). Auf dieser Basis wird in der folgenden Analyse die KS-Stelle unter die Lupe genommen. Ziel hierbei ist, herauszufinden, ob die Anwendung der jeweiligen KS-Regel mit Vorteilen verbunden war. Der Aufbau der Analyse ist wie folgt strukturiert:

5.4.1.1 Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regel

- Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

  Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

- Variablen: Summe der Fehler und Mittelwert der Fehleranzahl (von allen Fehlertypen) innerhalb der KS-Stelle; Variablentyp: ordinal

- Statistische Auswertung: Deskriptive Statistiken auf Basis der Fehlerannotation; Abbildungen: Balken

- Hypothesen:

  H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

  H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Die Variablen sind ordinal.

5.4.1.2 Zweiter Analysefaktor: Aufteilung der Annotationsgruppen

- In der Studie werden die Ergebnisse der Fehlerannotation in vier Annotationsgruppen unterteilt:

  (1) RR: MÜ ist vor und nach der Anwendung der KS-Regel fehlerfrei; (2) FF: MÜ beinhaltet vor und nach der Anwendung der KS-Regel Fehler; (3) RF:

MÜ ist nur vor der Anwendung der KS-Regel fehlerfrei; (4) FR: MÜ ist nur nach der Anwendung der KS-Regel fehlerfrei.

- Fragestellung: Wie hoch ist der Prozentsatz jeder Annotationsgruppe bei jeder Regel?

  Wie hoch ist der Prozentsatz jeder Annotationsgruppe *bei den einzelnen MÜ-Systemen* innerhalb der analysierten Regel?

- Statistische Auswertung: Häufigkeiten mit Bootstrapping[15] auf Basis der Fehlerannotation; Abbildungen: Kreisdiagramme

5.4.1.3 Dritter Analysefaktor: Vergleich der Fehlertypen vor vs. nach der Anwendung der KS-Regel

- Es werden die 13 analysierten Fehlertypen vor vs. nach der Anwendung der KS-Regel verglichen.

- Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

  Kommen bestimmte Fehlertypen *bei einem bestimmten MÜ-System* vor bzw. nach der Anwendung der KS-Regel vor?

  Davon wird abgeleitet, (1) ob bestimmte Fehlertypen, die vor der Anwendung der KS-Regel existierten, nach der Anwendung der KS-Regel eliminiert bzw. reduziert wurden; (2) ob bestimmte Fehlertypen erst nach der Anwendung der KS-Regel auftraten bzw. ihre Anzahl deutlich stieg (im Vergleich zu vor der Anwendung der KS-Regel).

- Variablen: Fehler existiert ja/nein; Fehlertyp: dichotom. Summe der Fehler sowie Mittelwert der Fehleranzahl innerhalb der KS-Stelle; Variablentyp: ordinal

- Statistische Auswertung: Kreuztabellen auf Basis der Fehlerannotation; Abbildungen: Balken

- Hypothesen:

  H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

[15] Bootstrapping ist ein statistisches Verfahren zur Schätzung der Stichprobenverteilung eines Schätzers durch erneute Stichprobenerstellung mit Ersatz aus der ursprünglichen Stichprobe. Es wird als ein nützliches Verfahren zum Testen der Modellstabilität erachtet. (IBM o.D.)

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

- Signifikanztest: McNemar; Begründung der Testauswahl: Mithilfe des McNemar-Tests können zwei verbundene dichotome Parameter verglichen werden, somit kann eine mögliche signifikante Veränderung bei einem Fehlertyp vor vs. nach der Anwendung der KS-Regeln identifiziert werden.

5.4.1.4 Vierter Analysefaktor: Vergleich der Qualität vor vs. nach der Anwendung der KS-Regel

- Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

  Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

- Variablen: Mittelwert der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch

- Statistische Auswertung: Deskriptive Statistiken auf Basis der Humanevaluation; Abbildungen: Fehlerbalken

- Hypothesen:

  H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

  H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

5.4.1.5 Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

- Fragestellung: Besteht bei der analysierten Regel ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS *minus* Qualität vor KS)?

Besteht *bei einem bestimmten MÜ-System* innerhalb der analysierten Regel ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS *minus* Qualität vor KS)?

- Variablen: Differenz der Mittelwerte der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch. Differenz der Fehleranzahl der einzelnen Fehlertypen; Variablentyp: ordinal.
- Statistische Auswertung: Spearman-Korrelationstest
- Hypothesen:

  H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

  H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

- Signifikanztest: Spearman-Korrelationstest; Begründung der Testauswahl: Eine der Variablen ist ordinal. Zudem setzt Spearman keine Anforderung an die Verteilung und die Linearität voraus.

5.4.1.6 Sechster Analysefaktor: Vergleich der Qualität vor vs. nach der Anwendung der KS-Regel auf Annotationsgruppenebene

- Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

  Davon wird abgeleitet, (1) ob bei der Gruppe RR die Stil- bzw. Inhaltsqualität vor bzw. nach der Anwendung der KS-Regel höher ist, obwohl die MÜ in beiden Fällen fehlerfrei ausfällt; (2) ob bei der Gruppe FF die Stil- bzw. Inhaltsqualität vor bzw. nach der Anwendung der KS-Regel höher ist, obwohl die MÜ in beiden Fällen Fehler beinhaltet; (3) ob bei der Gruppe RF die Stil- bzw. Inhaltsqualität nach der Anwendung der KS-Regel stieg, obwohl die MÜ nach der Anwendung der KS-Regel Fehler beinhaltet und davor fehlerfrei war; (4) ob bei der Gruppe FR die Stil- bzw. Inhaltsqualität nach der Anwendung der KS-Regel sank, obwohl die MÜ nach der Anwendung der KS-Regel fehlerfrei ist und davor Fehler beinhaltete.

- Variablen: Mittelwert der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala in jeder Annotationsgruppe; Variablentyp: metrisch
- Statistische Auswertung: Deskriptive Statistiken auf Basis der Humanevaluation unter Aufteilung der Daten nach den Annotationsgruppen; Abbildungen: Fehlerbalken
- Hypothesen:

  H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

  H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.
- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

5.4.1.7 Siebter Analysefaktor: Vergleich der AEM-Scores vor vs. nach der Anwendung der KS-Regel

- Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?
- Variablen: Mittelwert der AEM-Scores[16] sowie Differenzen der AEM-Scores (AEM-Score nach KS *minus* AEM-Score vor KS); Variablentyp: metrisch
- Statistische Auswertung: Deskriptive Statistiken auf Basis der automatischen Evaluation; Abbildungen: Fehlerbalken für die Differenzen der AEM-Scores[17]
- Hypothesen:

[16] Bei jeder MÜ wurde der Mittelwert der AEM-Scores auf Basis von zwei Referenzübersetzungen ermittelt; für eine genaue Beschreibung des Verfahrens siehe §4.5.6.4.

[17] Bei der Auswertung wurden nur die Differenzen der AEM-Scores (und nicht die Mittelwerte der AEM-Scores) verwendet. Der Grund ist, dass die Bewerter die komplette MÜ editiert haben. Ihre Edits können daher Stellen außerhalb der KS-Stelle umfassen. Da aber die MÜ vor und nach KS außerhalb der KS-Stelle vereinheitlicht wurden, wird hier davon ausgegangen, dass wir durch die Verwendung der Differenz (AEM-Score nach KS minus AEM-Score vor KS) nur die Edits innerhalb der KS-Stelle betrachten; für eine detaillierte Erläuterung siehe §4.5.6.4.

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

5.4.1.8 Achter Analysefaktor: Korrelation zwischen den AEM-Scores-Differenzen und der Qualitätsdifferenz

- Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores in TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS *minus* Mittelwert der AEM-Scores vor KS) und der Differenz der allgemeinen Qualität[18] (Qualität nach KS *minus* Qualität vor KS)?
- Variablen: Differenz der Mittelwerte der AEM-Scores sowie Differenz der Mittelwerte der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch
- Statistische Auswertung: Spearman-Korrelationstest
- Hypothesen:

  H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

  H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

- Signifikanztest: Spearman-Korrelationstest; Begründung der Testauswahl: Nicht alle Variablen sind normalverteilt. Spearman setzt keine Anforderungen an die Verteilung und die Linearität voraus.

[18] Die allgemeine Qualität ist der Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Untersuchung dieser Korrelation keine Unterscheidung zwischen der Stil- und Inhaltsqualität notwendig ist.

5.4.2 ERSTE REGEL: Für zitierte Oberflächentexte gerade Anführungszeichen "..." verwenden

5.4.2.1 Überblick

Im Folgenden wird die KS-Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ kurz beschrieben.[19] Zudem wird zusammenfassend und anhand eines Beispiels demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Für zitierte Oberflächentexte gerade Anführungszeichen verwenden (tekom-Regel-Nr. Z 103b)

Nach dieser Regel sollen Oberflächentexte, z. B. Texte in Software-Oberflächen oder Displaytexte von Geräten, in geraden Anführungszeichen stehen (tekom 2013: 117).

Begründung: Die Anführungszeichen erhöhen die Lesbarkeit. Im Vergleich zu der Verwendung von verschiedenen Schriftarten oder Schriftgraden sind gerade Anführungszeichen optisch nicht störend. Zudem unterstützen die Anführungszeichen eine korrekte Übersetzung. (ebd.: 118)

Umsetzungsmuster:

Vor KS: Oberflächentext ohne Anführungszeichen
Nach KS: Oberflächentext angegeben in geraden Anführungszeichen

KS-Stelle

Vor KS: Oberflächentext ohne Anführungszeichen
Nach KS: Oberflächentext mit geraden Anführungszeichen

Beispiele

Wählen Sie danach die Option *Software automatisch installieren.*

Wählen Sie danach die Option *"Software automatisch installieren"* .

Aufteilung der Testsätze: Die Länge der zitierten Oberflächentexte gemessen an der Anzahl der Wörter kann unterschiedlich sein. Da dieser Unterschied einen Einfluss auf das Ergebnis haben kann, wurde bei der Auswahl der Testsätze darauf geachtet, dass sie verschiedene Längen abdecken. Somit bestehen die 24 analysierten Sätze aus: 9 Sätzen mit nur einem Wort als

[19] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

Oberflächentext, 8 Sätzen mit zwei Wörtern und 7 Sätzen mit mehr als zwei Wörtern.

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.2.2 Vergleich der Fehleranzahl vor vs. nach der Verwendung von Anführungszeichen

Die Fehleranzahl sank deutlich um knapp 46 % von 137 Fehlern vor der Verwendung von Anführungszeichen (M = 1,14 / SD = ,964 / N = 120) auf 74 Fehler nach der Verwendung der Anführungszeichen (M = ,62 / SD = ,842 / N = 120), Abbildung 5.11 und Abbildung 5.12.[20] Der Mittelwert der Differenz (nach KS − vor KS) der Fehleranzahl pro Satz lag somit bei − ,53 (SD = ,767) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,67 (SD = ,653) und einem Maximum von − ,39 (SD = ,867) (Bootstrapping mit 1000 Stichproben). Die Differenz (nach KS − vor KS) der Fehleranzahl erwies sich als hochsignifikant (z (N = 120) = − 6,161 / p < ,001).

Durch die Kennzeichnung der Oberflächentexte mithilfe der Anführungszeichen konnten die MÜ-Systeme sie als spezifische Begriffe bzw. Mehrwortentitäten parsen und entsprechend richtig übersetzen: Die Testsätze beinhalteten 9 Sätze (von 24 Ausgangssätzen) mit einem Verb innerhalb des Oberflächentexts (z. B. ‚Upload vom Gerät', ‚Software automatisch installieren'). Bei 36 % (16 von 25) der Übersetzungen dieser Sätze wurden die Fehler behoben. Dieser Prozentsatz war höher im Vergleich zu den restlichen 15 Sätzen, in denen kein Verb im Oberflächentext vorkommt. In der letzteren Gruppe wurden in 28 % (21 von 75) der Übersetzungen die Fehler behoben. Sätze, die innerhalb des Oberflächentexts ein Verb enthalten, beinhalten zudem das Hauptverb des Satzes. Dies erschwert die syntaktische Analyse (Parsing) und begünstigt das Auftreten von Fehlern. In Tabelle 5.13 wurden die Fehler (Großschreibung und Wortstellung) in der Optionsbezeichnung nach der Verwendung der Anführungszeichen behoben (Genaueres zu den Fehlertypen unter §5.4.2.4).

[20] Abbildung 5.12 ist wie folgt zu lesen: Die Punkte zeigen, wie hoch die durchschnittliche Fehleranzahl ausfällt (z. B. 1,142 vor KS). Die Fehlerbalken zeigen das realisierte 95 %-Konfidenzintervall (CI) für die durchschnittliche Fehleranzahl (in dem Fall beläuft sich das CI vor KS auf ,98; 1,32). Demnach würde die Fehleranzahl bei der Durchführung einer weiteren vergleichbaren Untersuchung mit einer Wahrscheinlichkeit von 95 % zwischen ,98 und 1,32 liegen (vgl. Eckstein 2008: 81).

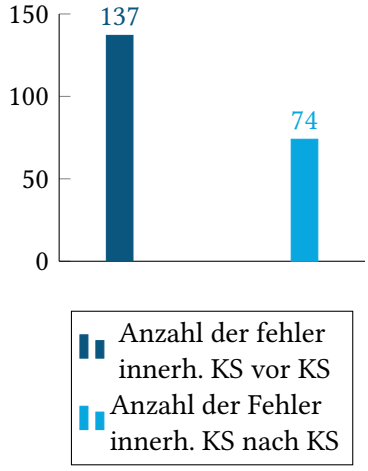


Abbildung 5.11: „Anführungsz. verw.“ - Fehlersumme vor vs. nach KS

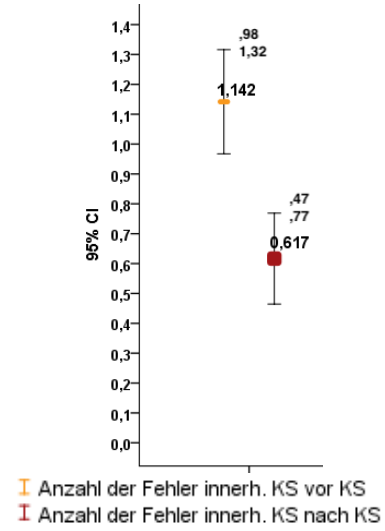


Abbildung 5.12: Anführungsz. verw.“ - Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

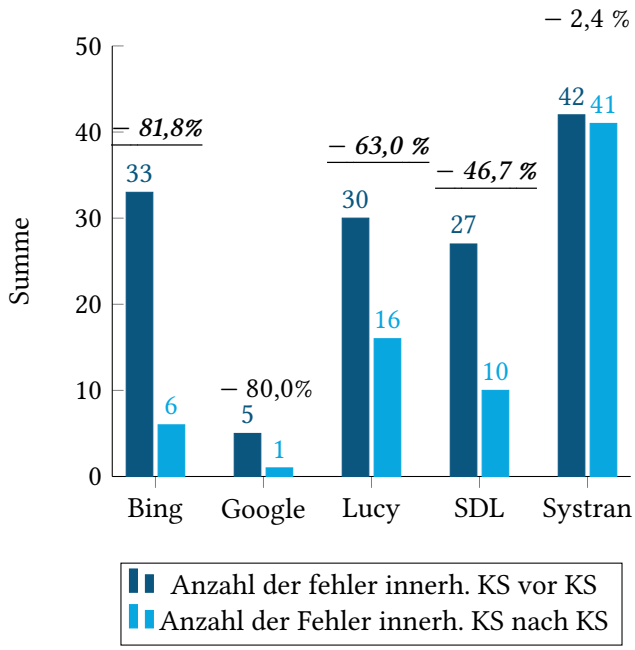
Tabelle 5.13: Beispiel 11

Vor-KS	Wählen Sie danach die Option **Software automatisch installieren**.
RBMÜ Lucy	Then select the option software automatically install.
Nach-KS	Wählen Sie danach die Option "**Software automatisch installieren**".
RBMÜ Lucy	Then select the option "Install software automatically".

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.2.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Durch die Kennzeichnung der Oberflächentexte mithilfe der Anführungszeichen konnten die MÜ-Systeme diese Texte als spezifische Begriffe bzw. Mehrwortentitäten parsen und entsprechend richtig übersetzen. Die Fehleranzahl nach der Umsetzung der KS-Regel sank entsprechend bei allen MÜ-Systemen.



Signifikante Differenz vor vs. nach KS

Abbildung 5.13: „Anführungsz. verw." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Bei dem NMÜ-System Google Translate war die Fehleranzahl sowohl vor als auch nach der Umsetzung der KS-Regel gering (Mdiff = – ,167), während bei dem HMÜ-System Systran die Fehleranzahl sowohl vor als auch nach der Umsetzung der KS-Regel sehr hoch und mit einer minimalen Differenz (nach KS – vor KS) verbunden war (Mdiff = – ,042). Ein Vergleich der MÜ beider Systeme (Google Translate vs. Systran) in Tabelle 5.14 zeigt Folgendes:

Bei dem HMÜ-System Systran wurde der Oberflächentext sowohl mit als auch ohne Verwendung der Anführungszeichen kleingeschrieben, jedoch gab es keinen Wortstellungsfehler. Das NMÜ-System konnte ohne Anführungszeichen den Oberflächentext nur zum Teil (in ‚Communication') als Bezeichnung erkennen

Tabelle 5.14: Beispiel 12

Vor-KS	Im Reiter **Kommunikation BACnet** können die notwendigen Einstellungen eingegeben werden.
HMÜ Systran	In the tab communication BACnet, the necessary settings can be entered.
GNMÜ	The necessary settings can be entered in the Communication tab BACnet.
Nach-KS	Im Reiter "**Kommunikation BACnet**" können die notwendigen Einstellungen eingegeben werden.
HMÜ Systran	In the tab “communication BACnet”, the necessary settings can be entered.
GNMÜ	The necessary settings can be entered in the "Communication BACnet" tab.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

und großschreiben. Problematisch für Google Translate war die unbekannte Bezeichnung ‚BACnet‘. Dies war möglicherweise die Ursache für den Wortstellungsfehler. Die Schwäche der NMÜ bei der Übersetzung von seltenen Wörtern und Eigennamen wurde in vorherigen Studien (vgl. Le & Schuster 2016; Koehn 2017) thematisiert. Die Verwendung der Anführungszeichen war hierbei für Google Translate nützlich, denn durch die Markierung der Oberflächenbezeichnung konnte der Wortstellungsfehler behoben werden. Insgesamt traten bei Google Translate fünf Fehler bei der Übersetzung von vier Sätzen (inkl. Tabelle 5.14) vor der Regelanwendung auf (Abbildung 5.13). In einer Wiederholung der Übersetzung Anfang 2020 ohne Verwendung der Anführungszeichen mit Google Translate trat nur ein Fehler (Kleinschreibung einer Bezeichnung) auf. Dies deutet auf einen Fortschritt im Bereich der Übersetzung von seltenen Wörtern bzw. Eigennamen hin.

Einen signifikanten Unterschied gab es bei den drei weiteren Systemen: dem *HMÜ-System Bing* (Mdiff = − 1,125; z (N = 24) = − 4,093 / p < ,001); dem *SMÜ-System SDL* (Mdiff = − ,708; z (N = 24) = − 3,117 / p = ,002); dem *RBMÜ-System Lucy* (Mdiff = − ,583; z (N = 24) = − 2,889 / p = ,004). Mehr Details zu den genauen Fehlertypen bei jedem System sind unter §5.4.2.4 zu finden.

5.4.2.3 Aufteilung der Annotationsgruppen

Wie Abbildung 5.14 zeigt, enthielten knapp 42 % der Übersetzungen sowohl vor als auch nach der Verwendung der Anführungszeichen Fehler (Gruppe FF). Außerdem waren fast 26 % der Übersetzungen sowohl vor als auch nach der Verwendung der Anführungszeichen fehlerfrei (Gruppe RR). Gleichzeitig wurden ungefähr 31 % falsche Übersetzungen (vor KS) nach der Verwendung der Anführungszeichen korrigiert (Gruppe FR).

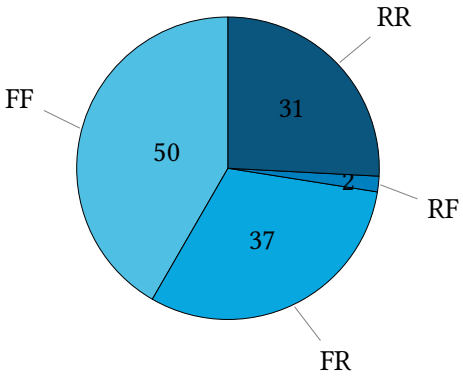


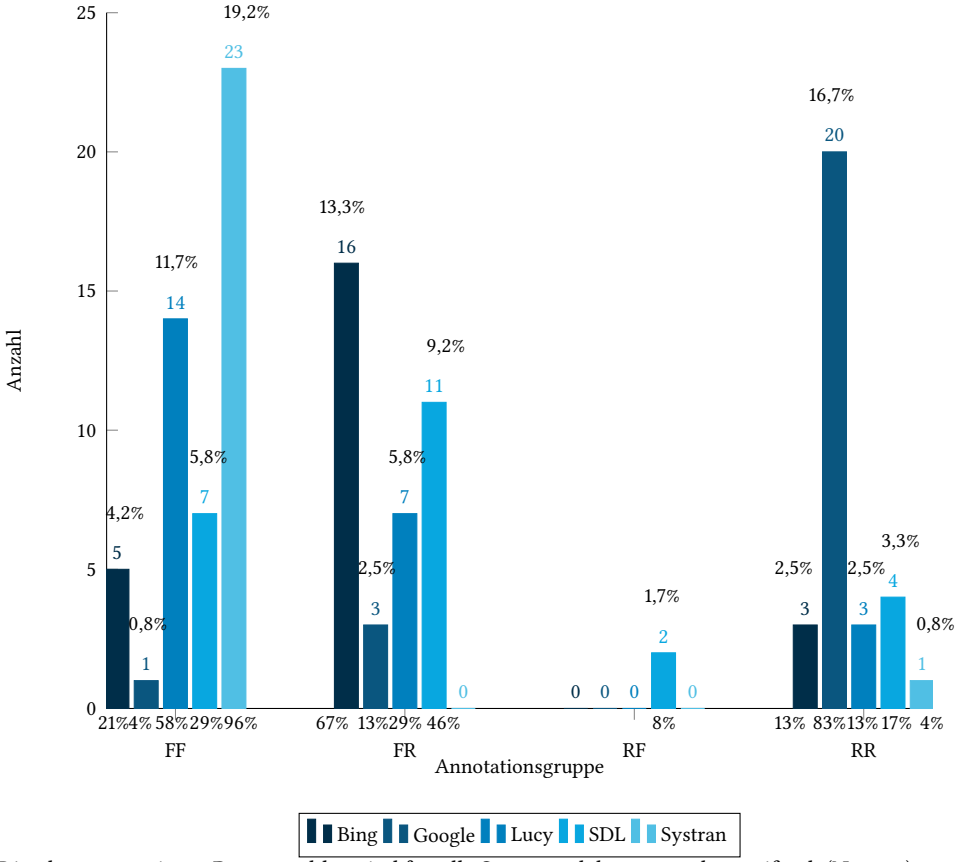
Abbildung 5.14: „Anführungsz. verw." – Aufteilung der Annotationsgruppen

Auf der anderen Seite kam es sehr selten vor (nur bei 2 Übersetzungen), dass nach der Verwendung der Anführungszeichen eine vor der Verwendung der Anführungszeichen korrekte Übersetzung falsch übersetzt wurde (Gruppe RF). Im nächsten Abschnitt (§5.4.2.4) werden die Fehlertypen ins Visier genommen.

5.4.2.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Eine genauere Analyse der einzelnen MÜ-Systeme bei der Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" zeigt Folgendes (Abbildung 5.15):

Wie Abbildung 5.15 zeigt, waren 96 % der 24 Sätze des HMÜ-Systems Systran sowohl vor der Verwendung der Anführungszeichen als auch danach falsch (Gruppe FF). Der Grund hierfür ist, dass Systran die Oberflächentexte immer in Kleinschreibung übersetzt hat (vgl. Tabelle 5.14). Ebenfalls blieben 11 von 24 Oberflächentexte bei dem RBMÜ-System Lucy trotz der Verwendung der Anführungszeichen kleingeschrieben (Gruppe FF). Insbesondere bei regelbasierten Systemen



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.15: „Anführungsz. verw.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen

ist davon auszugehen, dass sie auf Basis der hinterlegten Systemregeln die Oberflächentexte mithilfe der Anführungszeichen erkennen und in Großschreibung übersetzen. In anderen Worten würde bei dem RBMÜ-System in Verbindung mit dieser KS-Regel eine Besserung des MÜ-Outputs erwartet.

In der Gruppe RR war das NMÜ-System Google mit 83 % seiner 24 analysierten Sätze an erster Stelle vertreten. Zudem wurden 13 % seiner Sätze (3 von 24 Sätzen) erst nach der Verwendung der Anführungszeichen korrekt übersetzt (Gruppe FR) (vgl. Tabelle 5.14), siehe Abbildung 5.15.

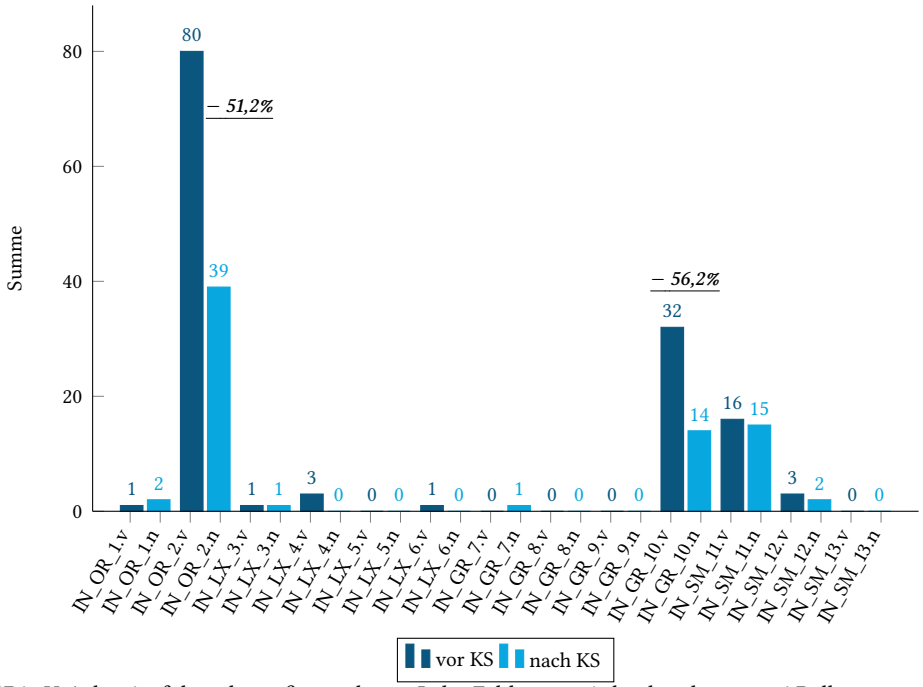
Im Gegenteil zu dem HMÜ-System Systran konnte das andere HMÜ-System (Bing) von der Anwendung der Regel deutlich profitieren, sodass 67 % der falschen Übersetzungen nach der Verwendung der Anführungszeichen fehlerfrei übersetzt wurden (Gruppe FR), siehe Abbildung 5.15. Auch bei dem SMÜ-System SDL konnten viele Kleinschreibungen mithilfe der Anführungszeichen (nach KS) korrigiert werden. Die Gruppe RF kam selten (genau in zwei Fällen) nur bei SDL vor.

5.4.2.4 Vergleich der Fehlertypen vor vs. nach der Verwendung von Anführungszeichen

Nach der Verwendung von Anführungszeichen sank die Fehleranzahl von zwei Fehlertypen deutlich: Fehlertyp OR.2 „Orthografie – Großschreibung“ von 80 auf 39 (51,2 % / Mv = ,67 / SDv = ,473 / Mn = ,33 / SDn = ,470 / N = 120) sowie Fehlertyp GR.10 „Grammatik – Falsche Wortstellung“ von 32 auf 14 (56,2 % / Mv = ,27 / SDv = ,463 / Mn = ,12 / SDn = ,322 / N = 120), siehe Abbildung 5.16. Der Unterschied bei den beiden Fehlertypen OR.2 und GR.10 erwies sich als hochsignifikant ($p < ,001$ / N = 120).

Die Markierung der Oberflächentexte, insbesondere derjenigen, die als Mehrwortentitäten auftreten, mithilfe von Anführungszeichen vereinfacht die syntaktische Analyse (Parsing). Entsprechend wird der Oberflächentext als solcher erkannt und großgeschrieben übersetzt. Ferner fördert die Erkennung des Oberflächentexts eine korrekte Wortstellung. Insbesondere Oberflächentexte, die ein Verb umfassen, sind für die MÜ-Systeme problematisch, da dieses Verb, wie in Tabelle 5.15, oft mit dem eigentlichen Verb des Satzes kollidiert:

Der Fehlertyp SM.11 „Semantik – Verwechslung des Sinns“ blieb fast unverändert (16 vor der Umsetzung bzw. 15 nach der Umsetzung der KS-Regel), siehe Abbildung 5.16. Wenn der Oberflächentext semantisch nicht eindeutig ist, stellt dies eine Schwierigkeit für die MÜ-Systeme dar, die durch die Verwendung von Anführungszeichen nicht gemindert werden kann. Beispiele hierfür sind die Oberflächentexte ‚Netzwerk *absuchen*‘ (wurde als ‚*Search* network‘ anstatt ‚Scan net-



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler vor KS und der zweite die Summe der Fehler nach KS, somit steht z. B. „OR_1.v“ für „OR_1: Orthografischer Fehler Nr. 1“ und „v: Vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).

Signifikante Differenz vor vs. nach KS

OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.16: „Anführungsz. verw.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

Tabelle 5.15: Beispiel 13

Vor-KS	**Hierzu kann die Funktion** Upload vom Gerät **gewählt werden.**
HMÜ Bing	To do this, the function can upload XXX the device be selected.
Nach-KS	**Hierzu kann die Funktion** "Upload vom Gerät" **gewählt werden.**
HMÜ Bing	To do this, the function "Upload from the unit" can be selected.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

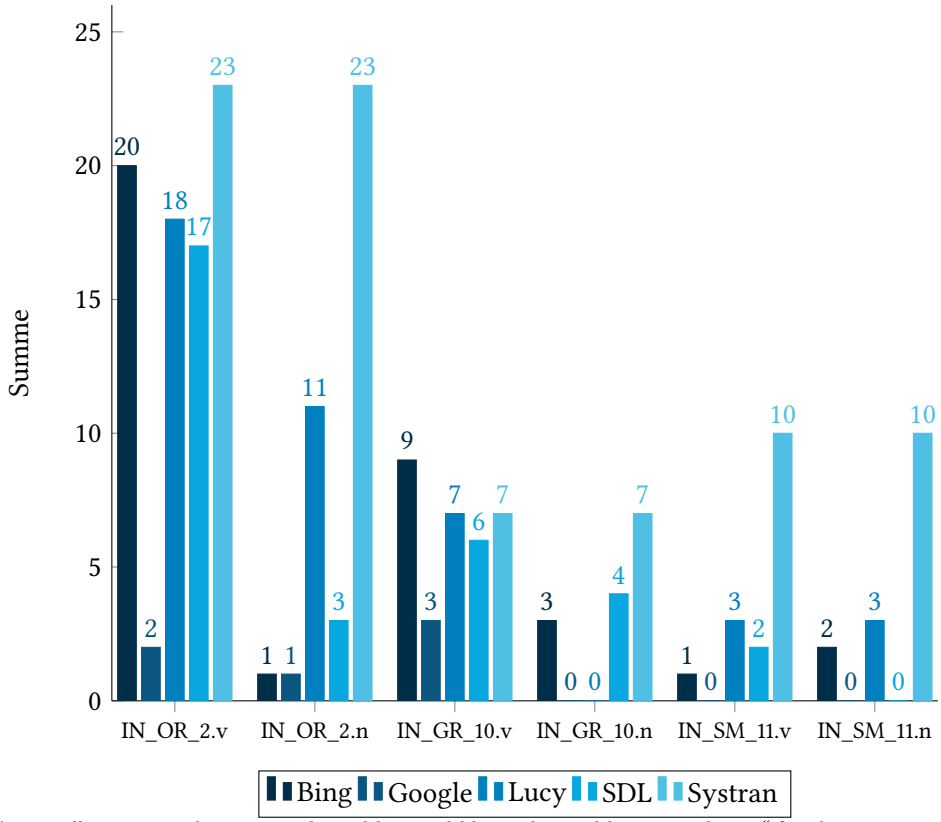
work‘ übersetzt) und ‚*Geräte*parameter ändern‘ (wurde als ‚Change *equipment* parameters‘ anstatt ‚Change device parameters‘ übersetzt). Alle anderen Fehlertypen kamen mit sehr wenigen Instanzen (max. drei Mal in den 240 analysierten maschinellen Übersetzungen) vor und nach der Anwendung der Regel vor.

5.4.2.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt, dass die obengenannten signifikanten Unterschiede in der Fehleranzahl des Fehlertyps OR.2 „Orthografie – Großschreibung“ und des Fehlertyps GR.10 „Grammatik – Falsche Wortstellung“ wie folgt zu beobachten sind: Die Veränderung im orthografischen Fehlertyp OR.2 „Großschreibung“ war bei den Systemen Bing, Lucy und SDL signifikant, während die Veränderung im grammatischen Fehlertyp GR.10 „Wortstellung“ nur bei Bing signifikant ausfiel.

Wie Abbildung 5.17 zeigt, sank der Fehlertyp OR.2 bei Bing von 20 auf 1 (– 95 %), bei Lucy von 18 auf 11 (– 38,9 %) und bei SDL von 17 auf 3 (– 82,4 %). Der Fehlertyp GR.10 sank bei Bing von 9 auf 3 (– 66,7 %). Tabelle 5.16 zeigt, wie die beiden Fehlertypen (der Großschreibungsfehler und der Wortstellungsfehler im Oberflächentext ‚Install software from a specific list‘) nach der Verwendung der Anführungszeichen behoben wurden:

Wie Abbildung 5.13 zeigt, fanden die markanten Veränderungen der Fehleranzahl nur bei den Systemen Bing, Lucy und SDL statt. Bei Google war die Fehleranzahl vor und nach der Anwendung der KS-Regel klein und bei Systran war die Fehleranzahl in beiden Szenarien groß mit einer minimalen Veränderung nach der Regelanwendung. Entsprechend erwies sich der Unterschied in der Fehleranzahl der beiden Fehlertypen OR.2 und GR.10 wie folgt als signifikant (Tabelle 5.17).



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen auftraten: In dieser Grafik kamen die Fehlertypen 5, 8, 9 und 13 bei gar keinem MÜ-System vor. Zudem wurden die Fehlertypen 1, 3, 4, 6, 7 und 12 nur einmal jeweils bei 1–3 MÜ-Systemen in vereinzelten Fällen registriert.

Abbildung 5.17: „Anführungsz. verw.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.16: Beispiel 14

Vor-KS	**Wählen Sie die Option** Software von einer bestimmten Liste installieren.
HMÜ Bing	Select the install option software from a specific list.
Nach-KS	**Wählen Sie die Option** "Software von einer bestimmten Liste installieren".
HMÜ Bing	Select the option "Install software from a specific list".

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Tabelle 5.17: „Anführungsz. verw." – Fehlertypen mit signifikanter Veränderung nach KS

	N	Mittelwert	Standard-abweichung	Signifikanz (McNemar-Test)
OR.2 „Großschreibung"				
Bing	24	vor KS = ,83 nach KS = ,04	vor KS = ,381 nach KS = ,204	p < ,001
Lucy	24	vor KS = ,75 nach KS = ,46	vor KS = ,442 nach KS = ,509	p = ,016
SDL	24	vor KS = ,71 nach KS = ,13	vor KS = ,464 nach KS = ,338	p < ,001
GR.10 „Falsche Wortstellung"				
Bing	24	vor KS = ,17 nach KS = ,13	vor KS = ,482 nach KS = ,338	p = ,031

Erwartungsgemäß blieb Fehlertyp SM.11 „Semantik – Verwechslung des Sinns“ fast unverändert. Wenn der Oberflächentext semantisch nicht eindeutig ist, stellt dies eine Schwierigkeit für die MÜ-Systeme dar, die durch die Verwendung von Anführungszeichen nicht beeinflusst werden können. Beispiel hierfür ist die Menübezeichnung ‚Raumdruck‘, die von Systran als ‚Space printing‘ und von Bing als ‚Pressure of space‘ anstelle von ‚Room pressure‘ übersetzt wurde.

5.4.2.5 Vergleich der MÜ-Qualität vor vs. nach der Verwendung von Anführungszeichen sowie die Korrelation zwischen den Fehlertypen und der Qualität

Sowohl die Stil- als auch die Inhaltsqualität[21] stiegen nach der Verwendung der Anführungszeichen stark: Die Stilqualität verbesserte sich um 13,6 % (Mv = 3,61 / SDv = ,654 / Mn = 4,10 / SDn = ,461 / N = 74). Die Inhaltsqualität erhöhte sich um 9,6 % (Mv = 4,17 / SDv = ,857 / Mn = 4,57 / SDn = ,527 / N = 74) (Abbildung 5.18). Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei ,490 (SD = ,525) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,368 und einem Maximum von ,611 und für die Inhaltsqualität bei ,394 (SD = ,763) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,217 und einem Maximum von ,571 (Bootstrapping mit 1000 Stichproben) (Abbildung 5.19). Die Differenzen (nach KS − vor KS) in der Stil- und Inhaltsqualität erwiesen sich als hochsignifikant ($z (N = 74) = -6{,}235 / p < ,001$) bzw. ($z (N = 74) = -4{,}740 / p < ,001$).

Die Humanevaluation deckt genauer auf, wie der Qualitätsanstieg zustande kam. Eine Analyse der Qualitätskriterien (Abbildung 5.20) zeigt, dass das erste und zweite Stilqualitätskriterium (SQ1 und SQ2) sowie das zweite Inhaltsqualitätskriterium (CQ2) die wesentliche Rolle bei der Qualitätsveränderung spielten.

In Tabelle 5.18 unterstützten die Anführungszeichen das MÜ-System (SDL) dabei, die Modulbezeichnung ‚ASV15‘ zu erkennen. Dies bewirkte eine Behebung des Wortstellungsfehlers.

Dieses Beispiel veranschaulicht, wie die Verwendung von Anführungszeichen bei Oberflächentexten zu einer korrekten Darstellung der MÜ (SQ1) führt. Dies wiederum trägt dazu bei, die Satzabsicht besser zu vermitteln (SQ2) und die Verständlichkeit zu erhöhen (CQ2).

[21] Definitionen der Qualität unter §4.5.5.1.

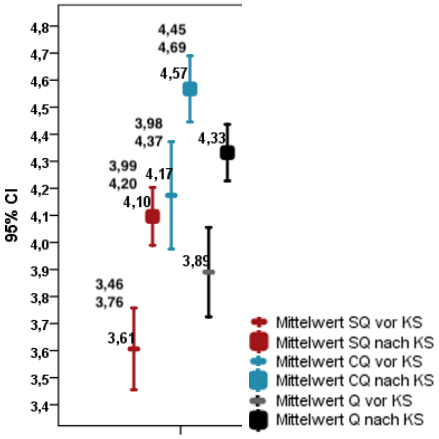


Abbildung 5.18: „Anführungsz. verw.“ – Mittelwerte der Qualität vor und nach KS

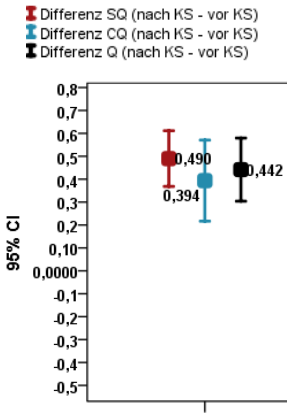
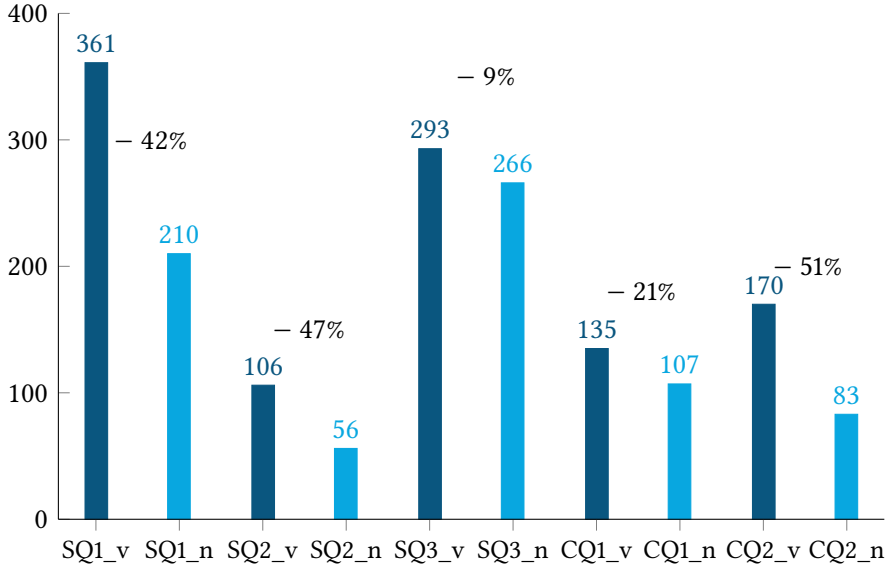


Abbildung 5.19: „Anführungsz. verw.“ – Mittelwert der Qualitätsdifferenzen

Tabelle 5.18: Beispiel 15

Vor-KS	Das Modul **ASV15** darf nur für seinen spezifizierten Einsatzzweck verwendet werden.
SMÜ SDL	The ASV module should only 15 be used for its specified operational purpose.
Nach-KS	Das Modul "**ASV15**" darf nur für seinen spezifizierten Einsatzzweck verwendet werden.
SMÜ SDL	The "ASV15" module should only be used for its specified operational purpose.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



SQ1: Ü ist **nicht** korrekt bzw. **nicht** klar dargestellt, d. h. nicht orthografisch
SQ2: Ü ist **nicht** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer **nicht** zum Handeln, zieht **nicht** seine Aufmerksamkeit an usw.
SQ3: Ü klingt **nicht** natürlich bzw. **nicht** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext **nicht** exakt wieder.
CQ2: Ü ist **nicht** leicht zu verstehen, d. h. **nicht** gut formuliert bzw. dargestellt.

Abbildung 5.20: „Anführungsz. verw." – Vergleich der Qualitätskriterien

5.4.2.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zeigte der Vergleich der Fehlertypen vor vs. nach der Verwendung der Anführungszeichen (siehe §5.4.2.4), dass ein Zusammenhang zwischen Fehlertyp OR.2 (Orthografie – Großschreibung) und Fehlertyp GR.10 (Grammatik – Falsche Wortstellung) und der Verwendung von Anführungszeichen besteht. Anhand der Spearman-Korrelationsanalyse erwies sich ein signifikanter starker negativer Zusammenhang zwischen der Differenz im Fehlertyp GR.10 (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Qualitätsdifferenz (Qualität nach KS – Qualität vor KS), siehe Tabelle 5.19. Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht nachgewiesen werden. Durch die Korrektur der Wortstellung nach der Verwendung der Anführungszeichen stieg die Qualität deutlich. In Tabelle 5.18 wurde Fehlertyp GR.10 eliminiert, daraufhin stiegen die Stilqualität um 1,00 Punkte und die Inhaltsqualität um 2,38 Punkte auf der Likert-Skala an.

Tabelle 5.19: „Anführungsz. verw.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl der **GR.10 „f. Wortstellung“**	74	< ,001	− ,532
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl der **GR.10 „f. Wortstellung“**	74	< ,001	− ,582
Differenz allg. Q (nach KS − vor KS)			
Diff. der Anzahl der **GR.10 „f. Wortstellung“**	74	< ,001	− ,593

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz ρ: Korrelationskoeffizient
schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

5.4.2.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.21 zeigt, stiegen sowohl die Stilqualität als auch die Inhaltsqualität bei allen MÜ-Systemen nach der Verwendung der Anführungszeichen.

Am höchsten stieg die Stilqualität des HMÜ-Systems Bing, nämlich um 21,2 %, und am wenigsten die des HMÜ-Systems Systran, und zwar um 9,7 %. Gleichzeitig stieg die Inhaltsqualität ebenfalls am stärksten bei Bing um 18,4 % und am schwächsten beim NMÜ-System Google Translate um knapp 3 %. Da im Falle von Google Translate 83 % der Sätze sowohl vor als auch nach der Verwendung der Anführungszeichen korrekt übersetzt wurden, zeigte sich bei der Inhaltsqualität kaum eine Veränderung. Dennoch verbesserte die Anwendung der Regel die Stilqualität signifikant um 9,4 %. Alle Differenzen bei den anderen MÜ-Systemen erwiesen sich in der Stil- und Inhaltsqualität als signifikant (Tabelle 5.20).

5.4.2.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Auf Basis der Fehlerannotation zeigte der Vergleich der Fehlertypen vor vs. nach der Verwendung der Anführungszeichen (siehe §5.4.2.4), dass ein Zusammen-

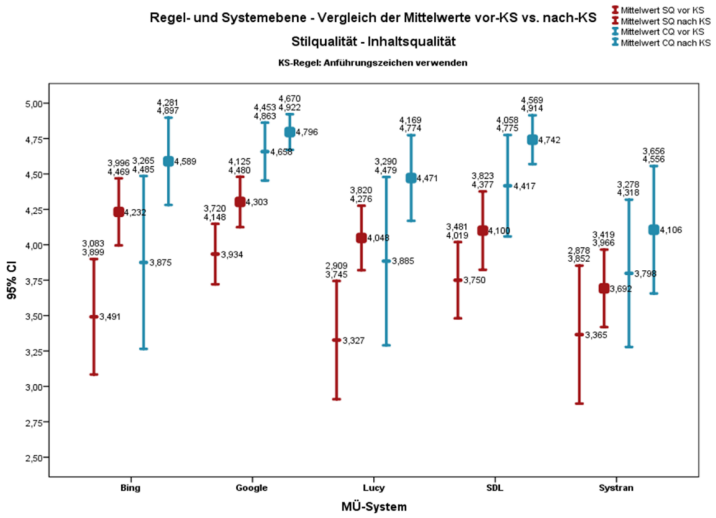


Abbildung 5.21: „Anführungsz. verw.“ – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.20: „Anführungsz. verw.“ – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	14	,003	− 2,944	14	,043	− 2,028	14	,016	− 2,418
Google	19	< ,001	− 3,520	19	,093	− 1,680	19	,001	− 3,250
Lucy	13	,005	− 2,812	13	,036	− 2,092	13	,004	− 2,908
SDL	15	,023	− 2,277	15	,018	− 2,374	15	,012	− 2,503
Systran	13	,031	− 2,163	13	,012	− 2,512	13	,011	− 2,554

p: Signifikanz z: Teststatistik nicht signifikant (p ≥ 0,05)

hang zwischen Fehlertyp OR.2 (Orthografie – Großschreibung) und Fehlertyp GR.10 (Grammatik – Falsche Wortstellung) und der Verwendung von Anführungszeichen besteht. Anhand der Spearman-Korrelationsanalyse erwies sich bei dem HMÜ-System Bing und dem RBMÜ-System Lucy jeweils ein signifikanter starker negativer Zusammenhang zwischen der Differenz in den Fehlertypen OR.2 sowie GR.10 und den Qualitätsdifferenzen. Durch die Korrektur der Großschreibung und der Wortstellung nach der Verwendung der Anführungszeichen stieg die Qualität. Weitere Korrelationen zwischen den Fehlertypen und der Qualität konnten bei den anderen MÜ-Systemen nicht nachgewiesen werden (Tabelle 5.21).

Tabelle 5.21: „Anführungsz. verw." – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Bing			Lucy		
	N	**p**	**ρ**	**N**	**p**	**ρ**
Differenz SQ (nach KS − vor KS)						
Diff. der Anzahl **OR.2 „Großs."**				13	,021	− ,631
Diff. der Anzahl **GR.10 „Wortst."**	14	,007	− ,686	13	,005	− ,722
Differenz CQ (nach KS − vor KS)						
Diff. der Anzahl **OR.2 „Großs."**				13	,142	− ,430
Diff. der Anzahl **GR.10 „Wortst."**	14	,016	− ,629	13	,001	− ,815
Differenz Q (nach KS − vor KS)						
Diff. der Anzahl **OR.2 „Großs."**				13	,126	− ,447
Diff. der Anzahl **GR.10 „Wortst."**	14	,005	− ,704	13	,001	− ,804

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >= 0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Der Zusammenhang lässt sich mithilfe von Beispielsätzen veranschaulichen: In Tabelle 5.16 wurden die Fehlertypen OR.2 und GR.10 eliminiert, daraufhin stiegen die Stilqualität (+ 1,50 Punkte) und die Inhaltsqualität (+ 2,0 Punkte auf der Likert-Skala) deutlich.

In Tabelle 5.22 wurde der Fehlertyp GR.10 eliminiert, daraufhin erfolgte ein starker Anstieg der Inhaltsqualität um 2,38 Punkte und der Stilqualität um 0,88 Punkte auf der Likert-Skala.

Tabelle 5.22: Beispiel 16

Vor-KS	Mit der Funktion **Anwendung neu konfigurieren**, kann dem Gerät eine neue Anwendung zugewiesen werden.
RBMÜ Lucy	Configure the function application again enables you to assign a new application to the device.
Nach-KS	Mit der Funktion "**Anwendung neu konfigurieren**", kann dem Gerät eine neue Anwendung zugewiesen werden.
RBMÜ Lucy	The function "Configure application again" enables you to assign a new application to the device.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Bei beiden Beispielen reflektieren die Kommentare der Bewerter in der Humanevaluation die Notwendigkeit der Großschreibung und der Hervorhebung der Funktionsbezeichnungen mithilfe von Anführungszeichen oder Fettformatierung. Ferner kommentierten sie aufgrund des Wortstellungsfehlers: „The sentence does not reflect the meaning of the source text at all"; „The translation fails in its informative function; the user has no idea what to do."[22]

5.4.2.6 Vergleich der MÜ-Qualität vor vs. nach der Verwendung von Anführungszeichen auf Annotationsgruppenebene

Die MÜ-Qualität[23] stieg in allen Annotationsgruppen mit Ausnahme der Gruppe RF (Abbildung 5.22):

[22] Außerdem haben die Bewerter im letzteren Beispiel eine Formulierung mit „Reconfigure" anstatt „Configure again" als idiomatische Übersetzung empfohlen. Da „Configure again" an sich keine semantisch falsche Übersetzung ist, wurde sie nicht als Fehler annotiert.

[23] Definitionen der Qualität unter §4.5.5.1.

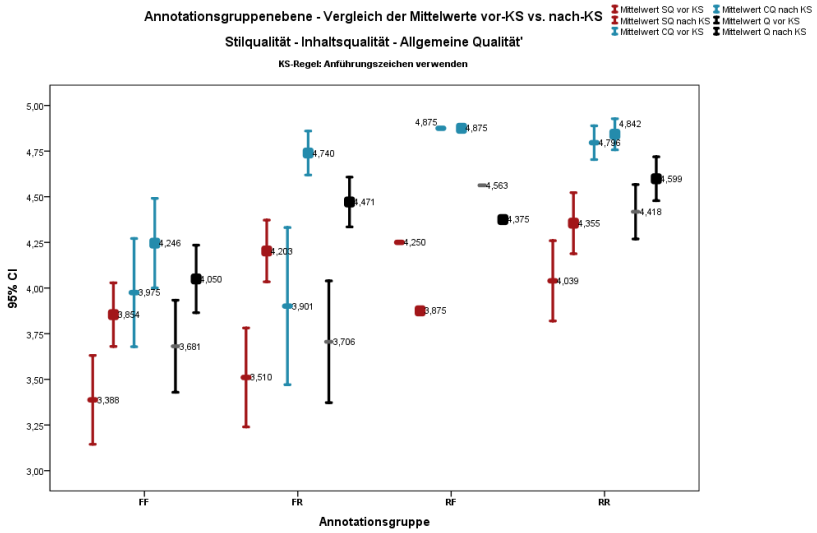


Abbildung 5.22: „Anführungsz. verw." – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

In der Gruppe FF (Übersetzung vor und nach der Verwendung der Anführungszeichen falsch) stieg die Qualität signifikant: bei der Stilqualität (z (N = 30) = − 3,711 / p < ,001) bzw. (z (N = 30) = − 2,998 / p = ,003) bei der Inhaltsqualität (Tabelle 5.23).

Dieses Ergebnis ist auf zwei Gründe zurückzuführen: Erstens wurden die Fehler, die vor der Verwendung der Anführungszeichen auftraten, nach der Verwendung der Anführungszeichen zum Teil eliminiert. In Tabelle 5.24 beinhaltet die Übersetzung einen orthografischen Fehler OR.2 „Großschreibung" und einen Wortstellungsfehler (GR.10) in der Optionsbezeichnung sowie einen semantischen Fehler SM.11 „Verwechslung des Sinns" (in der Übersetzung der Präposition ‚von') *vor* der Verwendung der Anführungszeichen und *nur* den semantischen Fehler *nach* der Verwendung der Anführungszeichen. Dies führte zur Steigerung der Stil- und Inhaltsqualität. Außerdem unterstützte die orthografische Darstellung mithilfe der Anführungszeichen eine schnelle Verständlichkeit, somit stieg die Stil- und Inhaltsqualität in Tabelle 5.24 um 1,63 bzw. 1,50 Punkte auf der Likert-Skala.

Zweitens gab es Fälle, in denen die Fehler sich vor sowie nach der Verwendung der Anführungszeichen wiederholten, dennoch stiegen die Stil- und Inhaltsqualität. In Tabelle 5.25 sind die Fehlertypen OR.2 (Großschreibung), SM.11

Tabelle 5.23: „Anführungsz. verw.“ – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS – vor KS)	30	< ,001	– 3,711
Differenz CQ (nach KS – vor KS)	30	,003	– 2,998
Differenz allg. Q (nach KS – vor KS)	30	< ,001	– 3,659
Annotationsgruppe FR			
Differenz SQ (nach KS – vor KS)	24	< ,001	– 3,950
Differenz CQ (nach KS – vor KS)	24	,001	– 3,402
Differenz allg. Q (nach KS – vor KS)	24	< ,001	– 3,787
Annotationsgruppe RF			
Differenz SQ (nach KS – vor KS)	1	–	–
Differenz CQ (nach KS – vor KS)	1	–	–
Differenz allg. Q (nach KS – vor KS)	1	–	–
Annotationsgruppe RR			
Differenz SQ (nach KS – vor KS)	19	,001	– 3,282
Differenz CQ (nach KS – vor KS)	19	,273	– 1,096
Differenz allg. Q (nach KS – vor KS)	19	,005	– 2,825

Tabelle 5.24: Beispiel 17

Vor-KS	Wählen Sie die Option **Software von einer bestimmten Liste installieren**.
RBMÜ Lucy	Select the option software of a certain list install.
Nach-KS	Wählen Sie die Option "**Software von einer bestimmten Liste installieren**".
RBMÜ Lucy	Choose the option "Install software of a certain list".

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

(Verwechslung des Sinnes) und GR.10 (Falsche Wortstellung) in den beiden Szenarien vorhanden. Eine Darstellung des Oberflächentexts in Anführungszeichen verbesserte jedoch die Stil- und Inhaltsqualität um 1,13 bzw. 0,88 Punkte auf der Likert-Skala.

Tabelle 5.25: Beispiel 18

Vor-KS	Wählen Sie die Option **Software von einer bestimmten Liste installieren**.
HMÜ Systran	Select the option software of a certain list install.
Nach-KS	Wählen Sie die Option "**Software von einer bestimmten Liste installieren**".
HMÜ Systran	Select the option “software of a certain list install”.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Erwartungsgemäß stieg die Qualität in der Gruppe FR (Übersetzung vor der Verwendung der Anführungszeichen falsch und nachher richtig) hochsignifikant bei der Stilqualität (z (N = 24) = − 3,950 / p < ,001) und bei der Inhaltsqualität (z (N = 24) = − 3,402 / p = ,001), siehe Tabelle 5.23. Die Verwendung der Anführungszeichen unterstützte die MÜ-Systeme bei der syntaktischen Analyse (Parsing), entsprechend wurden schwerwiegende Fehler wie die falsche Wortstellung eliminiert.

Die Gruppe RF (Übersetzung vor der Verwendung der Anführungszeichen richtig und nachher falsch) war, wie die Aufteilung der Annotationsgruppen (siehe §5.4.2.3) zeigte, sehr selten vertreten. Sie bestand aus nur zwei Übersetzungen. In dieser Gruppe blieb die Inhaltsqualität unverändert, während die Stilqualität sank. Aufgrund der kleinen Anzahl der Übersetzungen können sie allerdings als Ausnahmefälle betrachtet werden.

In der Gruppe RR (Übersetzung vor und nach der Verwendung der Anführungszeichen richtig) war der Unterschied in der Stilqualität signifikant (z (N = 19) = − 3,282 / p = ,001), wobei die Inhaltsqualität einen kleinen insignifikanten Anstieg zeigte. Dieses Ergebnis ist nachvollziehbar, denn die Übersetzung bei dieser Gruppe bleibt inhaltlich identisch, dennoch wird sie durch die Verwendung der Anführungszeichen klarer dargestellt. In Tabelle 5.26 gab es keine Differenz bei der Inhaltsqualität, wobei die Stilqualität um 0,63 Punkte auf der Likert-Skala stieg.

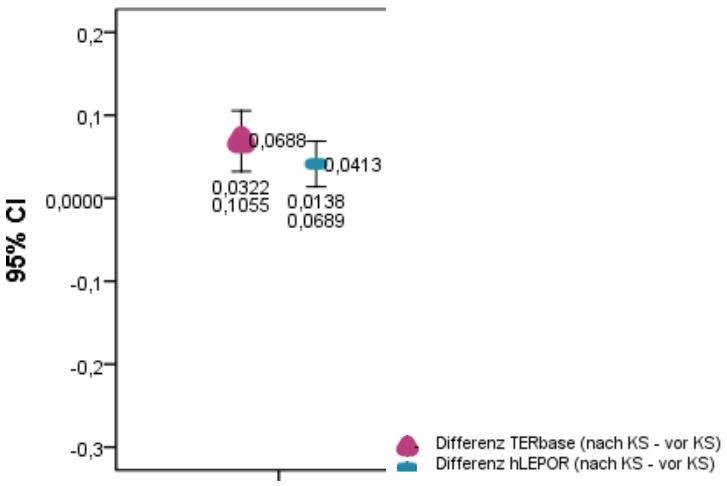
Tabelle 5.26: Beispiel 19

Vor-KS	Mit der Funktion **Geräteparameter ändern** können Sie die Parameter eines Gerätes ändern.
GNMÜ	The function Change device parameters enables you to change the parameters of a device.
Nach-KS	Mit der Funktion "**Geräteparameter ändern**" können Sie die Parameter eines Gerätes ändern.
GNMÜ	The function "Change device parameters" enables you to change the parameters of a device.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet.

5.4.2.7 Vergleich der AEM-Scores vor vs. nach der Verwendung von Anführungszeichen sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores vor und nach der Verwendung von Anführungszeichen zeigte sowohl mit TERbase als auch mit hLEPOR eine Verbesserung der AEM-Scores.



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.23: „Anführungsz. verw." – Mittelwert der Differenz der AEM-Scores

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei ,069 (SD = ,158) und für die hLEPOR bei ,041 (SD = ,119) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.23. Die Differenzen (nach KS − vor KS) in TERbase und hLEPOR erwiesen sich als signifikant (z (N = 74) = − 3,853 / p < ,001) bzw. (z (N = 74) = − 3,266 / p = ,001). Dieses Ergebnis deutet darauf hin, dass nach der Verwendung der Anführungszeichen weniger Edits erforderlich waren.

5.4.2.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter mittlerer positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität[24]. Tabelle 5.27 demonstriert die Korrelationswerte:

Tabelle 5.27: „Anführungsz. verw." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	74	< ,001	,431	mittlerer Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	74	,006	,319	mittlerer Zusammenhang

schwache Korrelation (ρ >= 0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Dieses Ergebnis weist darauf hin, dass – nach der Verwendung der Anführungszeichen – die Scores der beiden AEMs sich verbesserten und die Qualität stieg.

[24] Die allgemeine Qualität ist Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Untersuchung dieser Korrelation keine Unterscheidung zwischen der Stil- und Inhaltsqualität notwendig ist.

5.4.2.8 Analyse der ersten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der ersten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (–) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[25]

Regel 1: Für zitierte Oberflächentexte gerade Anführungszeichen "..." verwenden

Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Verwendung von Anführungszeichen

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene: H0 wurde abgelehnt und somit H1 bestätigt. Die Fehleranzahl sank nach der Verwendung der Anführungszeichen signifikant.	**Anz.F. (–)**
Auf Regel- und MÜ-Systemebene: Bei Lucy, SDL und Systran sank die Fehleranzahl nach der Verwendung der Anführungszeichen signifikant.	**Lu (–)** **SD (–)** **Sy (–)**
Bei Bing und Google sank ebenfalls nach KS die Fehleranzahl, dennoch war der Rückgang nicht signifikant.	**Bi (–)** **Go (–)**

[25] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Zweiter Analysefaktor

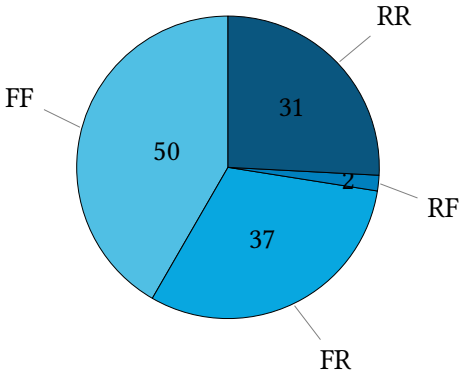


Abbildung 5.24: Aufteilung der Annotationsgruppen auf Regelebene

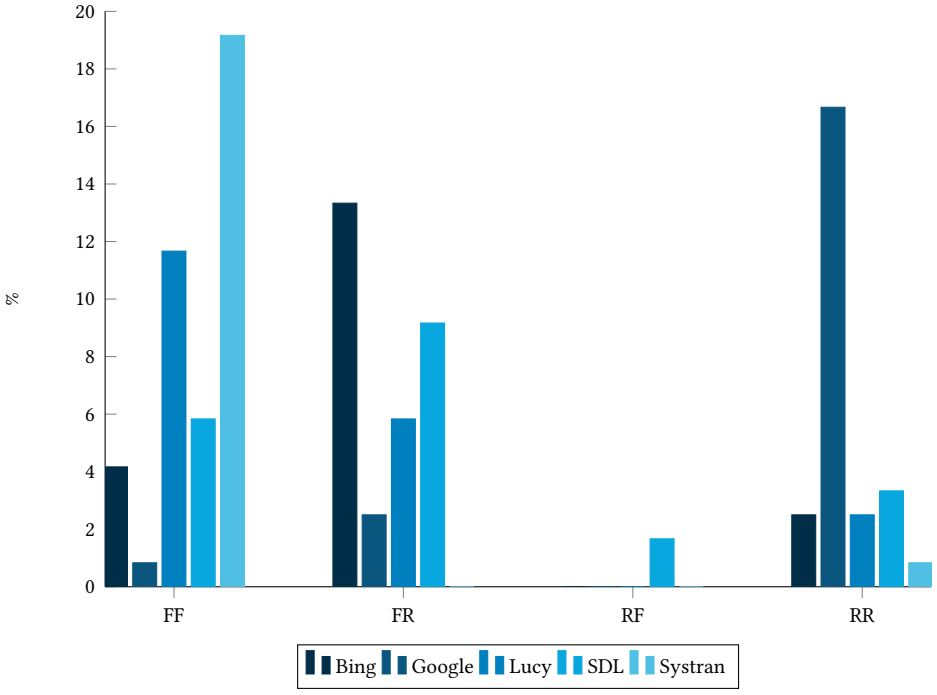


Abbildung 5.25: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Dritter Analysefaktor: Vergleich der Fehlertypen vor vs. nach der Verwendung von Anführungszeichen

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H1 wurde nur für zwei Fehlertypen bestätigt. **OR.2 (–)** **GR.10 (–)**
Die Fehleranzahl von OR.2 „Großschreibung“ und GR.10 „Falsche Wortstellung“ sanken nach der Verwendung der Anführungszeichen signifikant.

Auf Regel- und MÜ-Systemebene:
Bei Bing, Lucy und SDL sank die Fehleranzahl von OR.2 „Großschreibung“ nach der Verwendung der Anführungszeichen signifikant. **OR.2 (–): Bi Lu SD**

Bei Bing sank die Fehleranzahl von GR.10 „Falsche Wortstellung“ nach der Verwendung der Anführungszeichen signifikant. **GR.10 (–): Bi**

Alle weiteren Veränderungen waren nicht signifikant.

Vierter Analysefaktor: Vergleich der MÜ-Qualität vor vs. nach der Verwendung von Anführungszeichen

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Sowohl die Stil- als auch die Inhaltsqualität stiegen signifikant.

SQ (+)
CQ (+)

Auf Regel- und MÜ-Systemebene:
Die Stil- und Inhaltsqualität stiegen bei Bing, Lucy, SDL und Systran signifikant.

SQ (+)	CQ (+)
Bi Lu	**Bi Lu**
SD Sy	**SD Sy**

Bei Google stieg die Stilqualität signifikant und die Inhaltsqualität nicht signifikant.

SQ (+)	CQ (+)
Go	**Go**

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde nur für einen Fehlertyp bestätigt.
Es bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Falsche Wortstellung“ und der Differenz der Stil- und Inhaltsqualität.

***neg* GR.10 <<>> SQ**
***neg* GR.10 <<>> CQ**

Auf Regel- und MÜ-Systemebene:
Bei Bing bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Falsche Wortstellung“ und der Differenz der Stil- und Inhaltsqualität.

Bi
***neg* GR.10 <<>> SQ**
***neg* GR.10 <<>> CQ**

Bei Lucy bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Falsche Wortstellung“ und der Differenz der Stil- und Inhaltsqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des OR.2 „Großschreibung“ und der Differenz der Stilqualität.

Lu
***neg* GR.10 <<>> SQ**
***neg* GR.10 <<>> CQ**
***neg* OR.2 <<>> SQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität vor vs. nach der Verwendung von Anführungszeichen auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde nur zum Teil bestätigt:
Bei den Annotationsgruppen FF und FR stiegen die Stil- und Inhaltsqualität nach der Verwendung der Anführungszeichen signifikant.

SQ (+)
CQ (+)

Bei der Annotationsgruppe RR stieg die Stilqualität signifikant und die Inhaltsqualität nicht signifikant.

SQ (+)
CQ (+)

Bei der Annotationsgruppe RF sank die Stilqualität nicht signifikant und blieb die Inhaltsqualität unverändert.

SQ (–)
CQ (=)

Siebter Analysefaktor: Vergleich der AEM-Scores vor vs. nach der Verwendung von Anführungszeichen

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
AEM-Scores sowohl von TERbase als auch von hLEPOR verbesserten sich nach der Verwendung von Anführungszeichen signifikant.

TERbase (+)
hLEPOR (+)

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS – vor KS)

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.3 ZWEITE REGEL: Funktionsverbgefüge vermeiden

5.4.3.1 Überblick

Im Folgenden wird die KS-Regel „Funktionsverbgefüge vermeiden“ kurz beschrieben.[26] Zudem wird zusammenfassend und anhand eines Beispiels demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

[26] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

Beschreibung der KS-Regel: Funktionsverbgefüge vermeiden (tekom-Regel-Nr. L 103)

Nach dieser Regel soll das bedeutungstragende Verb anstatt des Funktionsverbgefüges verwendet werden (tekom 2013: 107).

Begründung: Die Verwendung des bedeutungstragenden Verbs lässt den Satz konkreter und direkter ausfallen (ebd.).

Umsetzungsmuster:

Vor KS: mit Funktionsverbgefüge
Nach KS: Das Funktionsverbgefüge wird durch das bedeutungstragende Verb ersetzt.

KS-Stelle

Vor KS: das Funktionsverbgefüge (FVG)
Nach KS: das bedeutungstragende Verb

Beispiele

Im oberen Abschnitt können Sie Einstellungen für die angezeigten Module vornehmen.

Im oberen Abschnitt können Sie die angezeigten Module einstellen.

Aufteilung der Testsätze: Der Datensatz besteht aus 24 verschiedenen Funktionsverbgefügen, die an unterschiedlichen Stellen in den Sätzen erscheinen und die zwei Formen eines Funktionsverbgefüges umfassen, nämlich:

12 Funktionsverbgefüge aus Funktionsverb + Präpositionalphrase (z. B. zur Anwendung kommen) und

12 Funktionsverbgefüge aus Funktionsverb + Nominalphrase (z. B. Auswahl treffen).

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.3.2 Vergleich der Fehleranzahl mit vs. ohne Funktionsverbgefüge

Die Fehleranzahl sank deutlich um 56,7 % von 90 Fehlern im Falle der Verwendung von Funktionsverbgefügen (M = ,75 / SD = ,928 / N = 120) auf 39 Fehler ohne Funktionsverbgefüge (M = ,33 / SD = ,552 / N = 120), Abbildung 5.26 und

Abbildung 5.27. Der Mittelwert der Differenz (nach KS – vor KS) der Fehleranzahl pro Satz lag somit bei – ,43 (SD =,984) mit einem 95%-Konfidenzintervall zwischen einem Minimum von – ,61 (SD = ,805) und einem Maximum von – ,26 (SD = 1,145) (Bootstrapping mit 1000 Stichproben). Die Differenz (nach KS – vor KS) der Fehleranzahl erwies sich als hochsignifikant (z (N = 120) = – 4,401 / p < ,001).

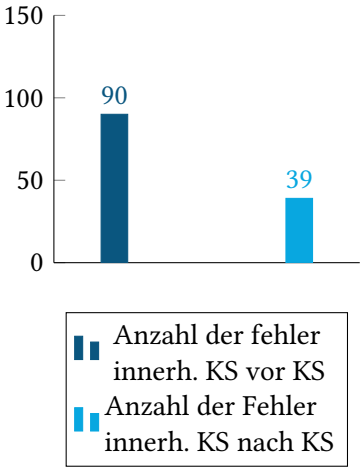


Abbildung 5.26: „FVG verm." – Fehlersumme vor vs. nach KS

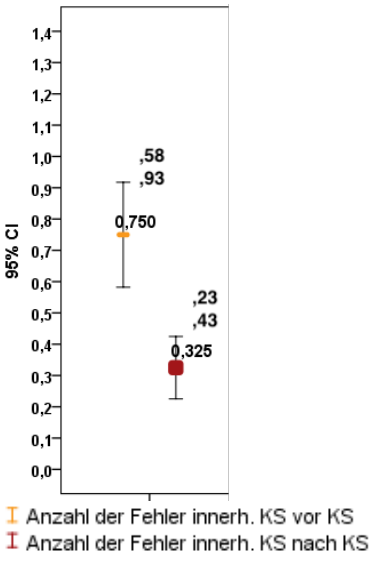


Abbildung 5.27: „FVG verm." – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Der Datensatz deckte gleichermaßen die zwei Formen eines Funktionsverbgefüges (12 aus Funktionsverb + Präpositionalphrase bestehende Funktionsverbgefüge (z. B. ‚zur Anwendung kommen') und 12 aus Funktionsverb + Nominalphrase bestehende Funktionsverbgefüge (z. B. ‚Auswahl treffen')) ab. Im Rahmen der Fehlerannotation wurde untersucht, ob eine der beiden Formen mit mehr Fehlern (vor der KS-Anwendung) verbunden war bzw. ob die KS-Anwendung bei einer der beiden Formen zur stärkeren Reduzierung der Fehleranzahl beitrug.

Wie Tabelle 5.28 zeigt, war der Rückgang der Fehleranzahl (– 3,75) im Falle der aus Funktionsverb + Präpositionalphrase bestehenden Funktionsverbgefüge im Vergleich zu den aus Funktionsverb + Nominalphrase bestehenden Funktionsverbgefügen (– 1,33) größer. Eine genauere Untersuchung des Datensatzes zeigte außerdem, dass die Anwendung der KS-Regel sich in zwei Fällen als besonders sinnvoll erwies:

Tabelle 5.28: Daten der untersuchten FVG-Formen

	FVG aus Funktionsverb + Präpositionalphrase	**FVG aus Funktionsverb + Nominalphrase**
Anzahl der Fälle	12 x 5 MÜ	12 x 5 MÜ
Durchschnittliche Diff. der Fehleranzahl (nach KS – vor KS)	− 3,75	− 1,33

Erstens – ein Funktionsverbgefüge besteht aus einem Substantiv und einem Funktionsverb. Wenn das Substantiv ein Teil eines Kompositums ist (z. B. Flecken*behandlung durchführen*, Küchen*montage durchführen*), stellt das Zerlegen dieses Kompositums (d. h. die Tokenisierung) und anschließend das Parsen des Substantives zusammen mit dem Funktionsverb als zusammengehörender Struktur für die älteren MÜ-Ansätze (SMÜ, RBMÜ und HMÜ Systeme) eine komplexe Aufgabe dar. In Tabelle 5.29 wurde das Kompositum ‚Küchenmöbelmontagen' bei der Verwendung des Funktionsverbgefüges vom HMÜ-System Bing gar nicht übersetzt bzw. vom RBMÜ-System Lucy falsch übersetzt. Nach der Verwendung des bedeutungstragenden Verbs (nach KS) lieferten beide Systeme korrekte Übersetzungen.

Zweitens – wenn das Funktionsverbgefüge kein Pendant bzw. keine direkte Übersetzung im Englischen hat (z. B. Einstellungen vornehmen, Schaden nehmen, Reinigung vornehmen), erwies sich die Verwendung des bedeutungstragenden Verbs ebenfalls als sinnvoll. In Tabelle 5.30 wurde der Kollokationsfehler in ‚make the cleaning' nach der Verwendung des bedeutungstragenden Verbs ‚reinigen' (nach KS) behoben.

Es gab insgesamt 19 falsche Übersetzungen, die eine dieser Eigenschaften aufweisen. Durch das Vermeiden des Funktionsverbgefüges wurden 13 von den 19 falschen Übersetzungen (68,4 %) korrigiert.

All diese problematischen Fälle des Funktionsverbgefüges (präpositionale FVG-Konstruktion, FVG in Verbindung mit Kompositum und FVG ohne englisches Pendant) gingen mit Transfer- und Parsing-Problemen einher. Somit wurde das Funktionsverbgefüge wörtlich übersetzt, was zu lexikalischen und semantischen Fehlern führte (wie in der Analyse der Fehlertypen und ihrer Korrelationen mit den Qualitätsveränderungen unter §5.4.3.4 und §5.4.3.5 detailliert dargestellt).

Tabelle 5.29: Beispiel 20

Vor-KS	Küchenmöbel**montagen** dürfen nur von geschulten Fachleuten **durchgeführt** werden.
HMÜ Bing	Küchenmöbel**montagen** may only be carried out by trained specialists.
RBMÜ Lucy	Kitchen piece of furniture assemblies may only be carried out by trained specialists.
Nach-KS	Küchenmöbel dürfen nur von geschulten Fachleuten **montiert** werden.
HMÜ Bing	Kitchen furniture may only be installed by trained specialists.
RBMÜ Lucy	Kitchen furniture may only be mounted by trained specialists.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Tabelle 5.30: Beispiel 21

Vor-KS	Die **Reinigung** der Küchenmöbel sollten Sie mit einem leicht feuchten Tuch **vornehmen**.
HMÜ Bing	You should make the cleaning of the kitchen furniture with a slightly damp cloth.
Nach-KS	Sie sollten die Küchenmöbel mit einem leicht feuchten Tuch **reinigen**.
HMÜ Bing	You should clean the kitchen furniture with a slightly damp cloth.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.3.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Durch die Verwendung des bedeutungstragenden Verbs (nach KS) wurden viele Kollokationsfehler vermieden, die bei der Verwendung des Funktionsverbgefüges (vor KS) auftraten. Die Fehleranzahl nach der Umsetzung der KS-Regel sank bei allen Systemen:

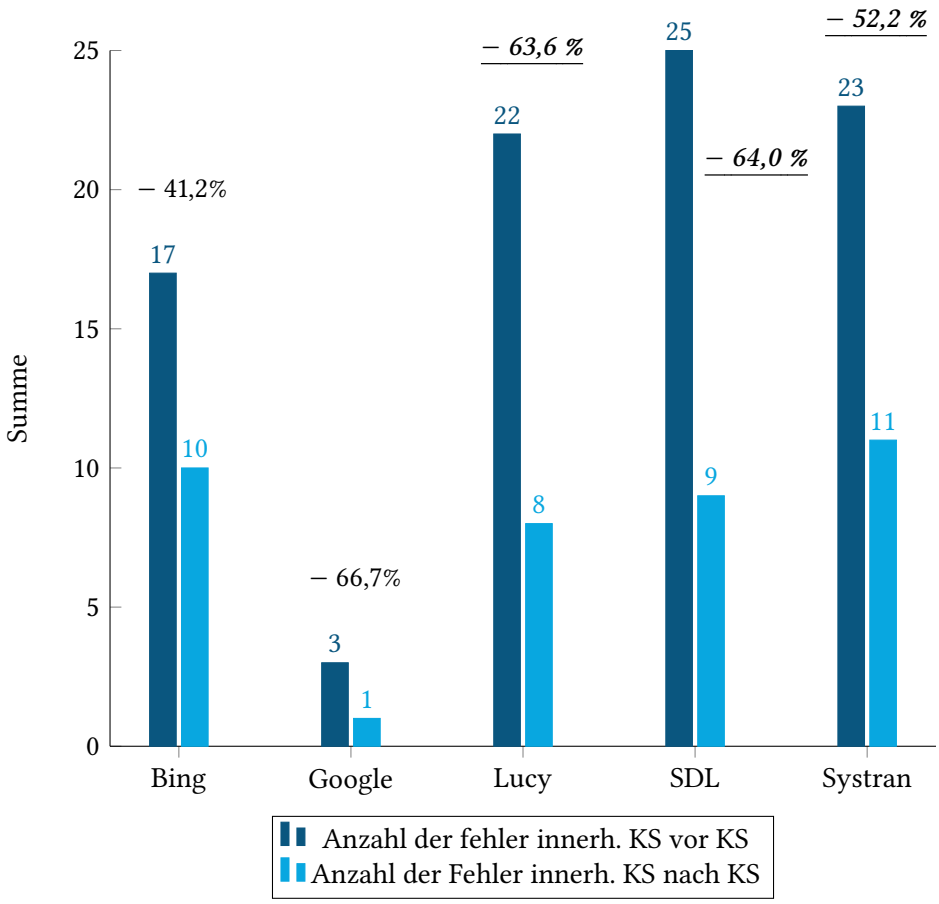


Abbildung 5.28: „FVG verm." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Signifikant war der Rückgang bei dem RBMÜ-System Lucy (Mdiff = – ,583; z (N = 24) = – 2,501 / p = ,012); dem SMÜ-System SDL (Mdiff = – ,667; z (N = 24) = – 2,358 / p = ,018); dem HMÜ-System Systran (Mdiff = – ,500; z (N = 24) =

– 2,144 / p = ,032). Bei dem NMÜ-System Google war die Fehleranzahl sowohl vor als auch nach der Umsetzung der KS-Regel gering (Mdiff = – ,083). Bei dem HMÜ-System Bing war die Differenz (nach KS – vor KS) sehr niedrig (Mdiff = – ,292).

Das NMÜ-System Google Translate war in der Lage, die Sätze mit und ohne Funktionsverbgefüge korrekt – und in mehreren Fällen sogar identisch – zu übersetzen, während die Übersetzungen der anderen MÜ-Systeme Fehler beinhalteten. In Tabelle 5.30 konnte das HMÜ-System Bing erst nach der Verwendung des bedeutungstragenden Verbs (nach KS) den Satz korrekt übersetzen. Das SMÜ-System SDL konnte denselben Satz weder mit Funktionsverbgefüge noch mit dem bedeutungstragenden Verb korrekt übersetzen (die Fehlertypen werden unter §5.4.3.4 diskutiert), während die Übersetzung vom NMÜ-System Google Translate in beiden Fällen trotz der veränderten Struktur des Ausgangssatzes korrekt und identisch war (Tabelle 5.31).

Tabelle 5.31: Beispiel 22

Vor-KS	Die **Reinigung** der Küchenmöbel sollten Sie mit einem leicht feuchten Tuch **vornehmen**.
SMÜ SDL	The cleaning of the kitchen furniture you should start with a slightly damp cloth.
GNMÜ	You should clean the kitchen furniture with a slightly damp cloth.
Nach-KS	Sie sollten die Küchenmöbel mit einem leicht feuchten Tuch **reinigen**.
SMÜ SDL	You should set the kitchen furniture with a slightly damp cloth.
GNMÜ	You should clean the kitchen furniture with a slightly damp cloth.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Sogar ein gängiges Funktionsverbgefüge wie ‚zur Verfügung stehen' war für alle MÜ-Systeme problematisch zu übersetzen, während das NMÜ-System es problemlos übersetzen konnte und erneut sowohl mit dem Funktionsverbgefüge als auch mit dem bedeutungstragenden Verb eine identische Übersetzung lieferte. Tabelle 5.32 veranschaulicht, wie die Übersetzung des RBMÜ-Systems Lucy vor und nach der Anwendung der KS-Regel Fehler beinhaltete, während die Übersetzung des NMÜ-Systems Google Translate in den beiden Szenarien fehlerfrei war.

Tabelle 5.32: Beispiel 23

Vor-KS	Auf der Startseite **stehen** die folgenden Funktionen zur Auswahl **zur Verfügung**.
RBMÜ Lucy	On the Start page, the following functions are to choose from availably.
GNMÜ	The following functions are available for selection on the start page.
Nach-KS	Auf der Startseite **sind** die folgenden Funktionen zur Auswahl **vorhanden**.
RBMÜ Lucy	On the Start page, the following functions to choose from are available.
GNMÜ	The following functions are available for selection on the start page.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.3.3 Aufteilung der Annotationsgruppen

Knapp 42 % der Übersetzungen waren sowohl mit Funktionsverbgefüge als auch mit dem bedeutungstragenden Verb richtig (Gruppe RR). Außerdem waren fast 21 % der Übersetzungen in beiden Szenarien falsch (Gruppe FF). Gleichzeitig wurden ungefähr 30 % falsche Übersetzungen von Sätzen mit Funktionsverbgefügen nach der Verwendung des bedeutungstragenden Verbs korrigiert (Gruppe FR) (siehe Abbildung 5.29). Auf der anderen Seite gab es 10 Fälle (8,3 %), die nur nach der Verwendung des bedeutungstragenden Verbs (nach KS) falsch übersetzt wurden (Gruppe RF) (siehe Abbildung 5.29).

Hierbei handelte es sich zum Teil um semantische Fehler und in einzelnen Fällen um grammatische Fehler. Aufgrund der kleinen Anzahl der Fälle dieser Gruppe konnte jedoch kein bestimmtes Fehlermuster erkannt werden. Im nächsten Abschnitt werden die Fehlertypen genauer besprochen.

5.4.3.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Eine genauere Analyse der einzelnen MÜ-Systeme bei der Regel „Funktionsverbgefüge vermeiden“ zeigt Folgendes:

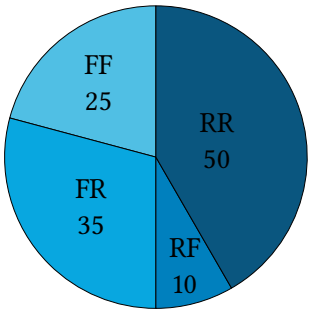
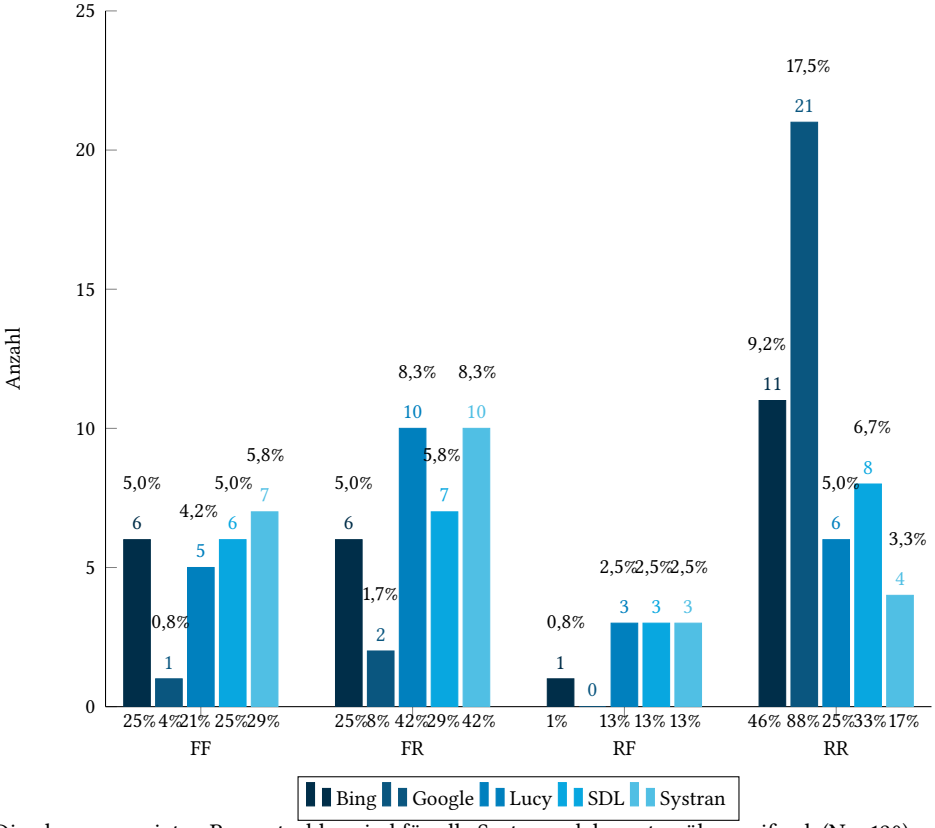


Abbildung 5.29: „FVG verm.“ – Aufteilung der Annotationsgruppen



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.30: „FVG verm.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen

Die größten Prozentsätze von drei Systemen wurden bei der Gruppe RR verzeichnet. Diese sind das NMÜ-System Google mit 88 %, das HMÜ-System Bing mit 46 % sowie das SMÜ-System SDL mit 33 % (Abbildung 5.30). Die zwei weiteren Systeme, das RBMÜ-System Lucy und das HMÜ-System Systran, waren stärker repräsentiert bei der Gruppe FR (im Vergleich zur Gruppe RR), jeweils mit 42 % (Abbildung 5.30). Durch das Vermeiden von den Funktionsverbgefügen konnten sämtliche Kollokationsfehler behoben werden. Dennoch blieb die Gruppe FF bei allen Systemen mit Ausnahme von Google relativ groß (mehr als 20 %) (Abbildung 5.30). Im folgenden Abschnitt werden die aufgetreten Fehlertypen ins Visier genommen.

5.4.3.4 Vergleich der Fehlertypen mit vs. ohne Funktionsverbgefüge

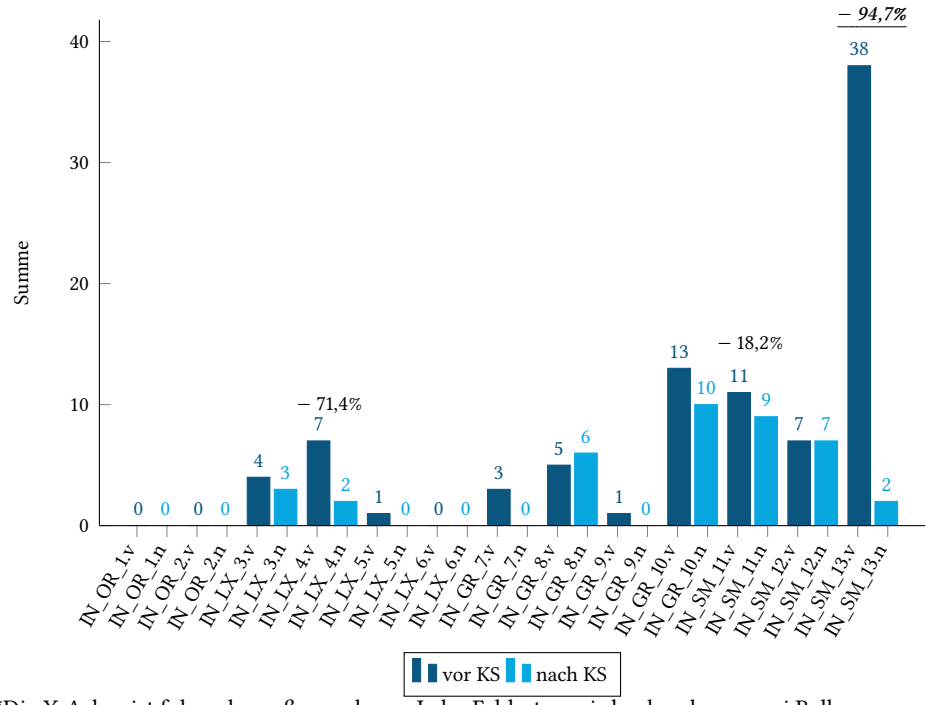
Nach der Formulierung des Satzes ohne Funktionsverbgefüge sank die Fehleranzahl bei dem Fehlertyp SM.13 „Semantik – Kollokationsfehler“ deutlich, und zwar von 38 auf 2 (– 94,7 % / Mv = ,32 / SDv = ,502 / Mn = ,02 / SDn = ,129 / N = 120) (Abbildung 5.31). Der Unterschied in der Fehleranzahl erwies sich als hochsignifikant (p < ,001 / N = 120).

Erwartungsgemäß ging die Anzahl der Kollokationsfehler nach der Anwendung der KS-Regel zurück, denn nicht alle deutschen Funktionsverbgefüge haben ein Pendant im Englischen. Ein Beispiel hierfür ist das Funktionsverbgefüge ‚in Betrieb nehmen‘ (siehe Tabelle 5.33). Hier konnte das MÜ-System nur das bedeutungstragende Verb (nach KS) richtig übersetzen.

Tabelle 5.33: Beispiel 24

Vor-KS	Der Bediener darf erst die Maschine **in Betrieb nehmen**, wenn er die Betriebsanleitung gelesen hat.
HMÜ Systran	The operator may only take the machine in enterprise after reading the operating instructions.
Nach-KS	Der Bediener darf erst die Maschine **starten**, wenn er die Betriebsanleitung gelesen hat.
HMÜ Systran	The operator may only start the machine after reading the operating instructions.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler *vor KS* und der zweite die Summe der Fehler *nach KS*, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
__*Signifikante Differenz vor vs. nach KS***__
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.31: „FVG verm.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

Bei allen anderen Fehlertypen hat sich die Fehleranzahl vor im Vergleich zu nach der Anwendung der Regel nicht (signifikant) verändert, z. B. sank der semantische Fehlertyp SM.11 „Verwechslung des Sinns“ von 11 auf 9 (– 18,2 %) Fehler und der lexikalische Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ sank von 7 auf 2 Fehler (– 71,4 %; ein aufgrund der kleinen Fehleranzahl nicht signifikanter Rückgang), während der semantische Fehlertyp SM.12 „Falsche Wahl“ unverändert blieb (7 Fehler vor und nach KS).

5.4.3.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt (Abbildung 5.32), dass Fehlertyp SM.13 „Semantik – Kollokationsfehler“ bei zwei Systemen sank (RBMÜ-System Lucy und HMÜ-System Systran) und bei drei Systemen vollständig behoben wurde (HMÜ- System Bing, NMÜ-System Google und SMÜ-System SDL).

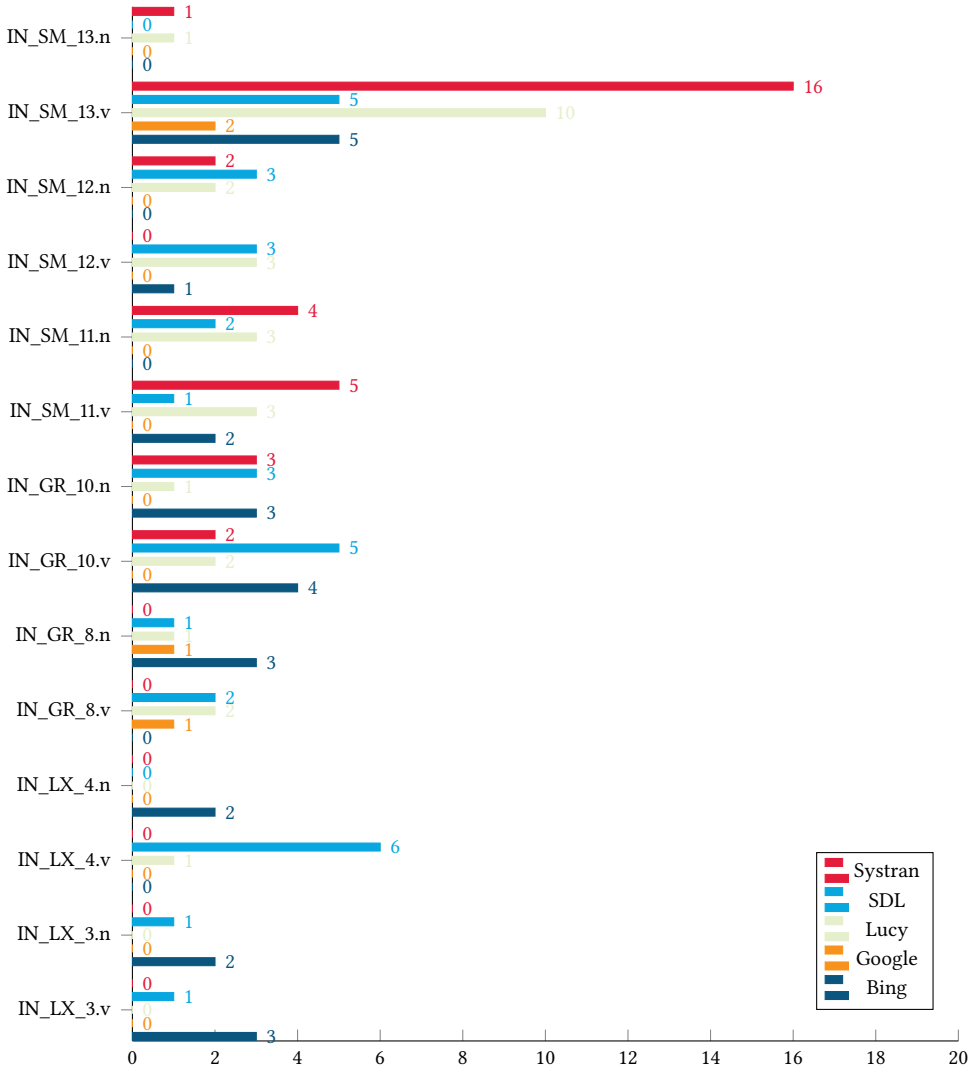
Nur bei dem RBMÜ-System Lucy (ein Rückgang von 10 auf 1 Fehler (– 90 %)) und dem HMÜ-System Systran (ein Rückgang von 16 auf 1 Fehler (– 93,8 %)) erwies sich die Differenz bei dem SM.13 Kollokationsfehler als signifikant.

Tabelle 5.34: „FVG verm.“ – Fehlertypen mit signifikanter Veränderung nach KS

	N	Mittelwert	Standard-abweichung	Signifikanz (McNemar-Test)
SM.13 „Kollokationsfehler“				
Lucy	24	vor KS = ,42 nach KS = ,04	vor KS = ,504 nach KS = ,204	p = ,004
Systran	24	vor KS = ,67 nach KS = ,04	vor KS = ,637 nach KS = ,204	p < ,001

Tabelle 5.35 zeigt den bei dem RBMÜ-System Lucy aufgetretenen Kollokationsfehler im Falle der Verwendung des Funktionsverbgefüges und wie er nach der Verwendung des bedeutungstragenden Verbs vermieden wurde.

Wie Abbildung 5.32 zeigt, gab es bei dem zweiten HMÜ-System (Bing) sowie bei dem SMÜ-System SDL bei mehreren Fehlertypen eine kleine Fehleranzahl. Schließlich war bei dem NMÜ-System Google Translate die Fehleranzahl im Allgemeinen sehr gering (insgesamt 3 Fehler vor KS und 1 Fehler nach KS), wie die Analyse der Fehleranzahl (Abbildung 5.28) zeigt.



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 1, 2 und 6 bei gar keinem MÜ-System vor. Zudem kamen die Fehlertypen 5, 7 und 9 nur einmal jeweils bei 1 bis 3 MÜ-Systemen in vereinzelten Fällen vor.

Abbildung 5.32: „FVG verm.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.35: Beispiel 25

Vor-KS	Wird diese Regel nicht beachtet, kann der Motor **Schaden nehmen**.
RBMÜ Lucy	If this rule is not observed, the motor can take damage.
Nach-KS	Wird diese Regel nicht beachtet, kann der Motor **beschädigt werden**.
RBMÜ Lucy	If this rule is not observed, the motor can be damaged.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.3.5 Vergleich der MÜ-Qualität mit vs. ohne Funktionsverbgefüge sowie die Korrelation zwischen den Fehlertypen und der Qualität

Sowohl die Stil- als auch die Inhaltsqualität[27] stiegen nach der Vermeidung von Funktionsverbgefügen. Auf den ersten Blick erkennt man in Abbildung 5.34, dass der Einfluss auf die Stilqualität größer als der auf die Inhaltsqualität war. Die Übersetzung des Funktionsverbgefüges war in vielen Fällen mit dem semantischen Fehlertyp SM.13 „Kollokationsfehler“ verbunden (siehe §5.4.3.4). Nachdem das Funktionsverbgefüge vermieden wurde (nach KS), sank die Anzahl der Kollokationsfehler stark. Dies trug zu einer besseren Verständlichkeit (höhere Inhaltsqualität) und gleichzeitig zu einer deutlich natürlicheren Formulierung (höhere Stilqualität) bei.

Die Stilqualität verbesserte sich um 8,2 % (Mv = 4,03 / SDv = ,583 / Mn = 4,36 / SDn = ,449 / N = 84). Die Inhaltsqualität erhöhte sich um 5,2 % (Mv = 4,25 / SDv = ,792 / Mn = 4,47 / SDn = ,681 / N = 84), siehe Abbildung 5.33. Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei ,330 (SD = ,615) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,197 und einem Maximum von ,464 und für die Inhaltsqualität bei ,219 (SD = ,948) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,013 und einem Maximum von ,425 (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.34. Die Differenzen (nach KS − vor KS) in der Stil- und Inhaltsqualität erwiesen sich als signifikant (z (N = 84) = − 4,036 / p < ,001) bzw. (z (N = 84) = − 2,180 / p = ,029).

Wie unter §5.4.3.2 dargestellt, waren die Funktionsverbgefüge der Gruppe Funktionsverb + Präpositionalphrase mit einem größeren Rückgang der Fehler-

[27] Definitionen der Qualität unter §4.5.5.1.

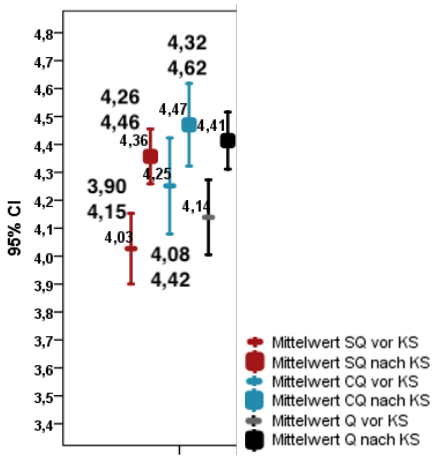


Abbildung 5.33: „FVG verm." – Mittelwerte der Qualität vor und nach KS

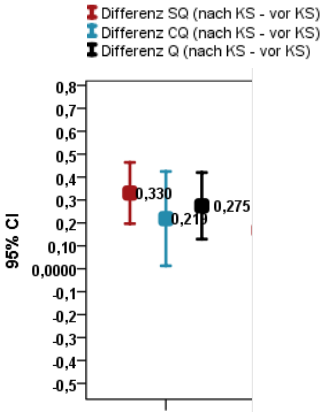


Abbildung 5.34: „FVG verm." – Mittelwert der Qualitätsdifferenzen

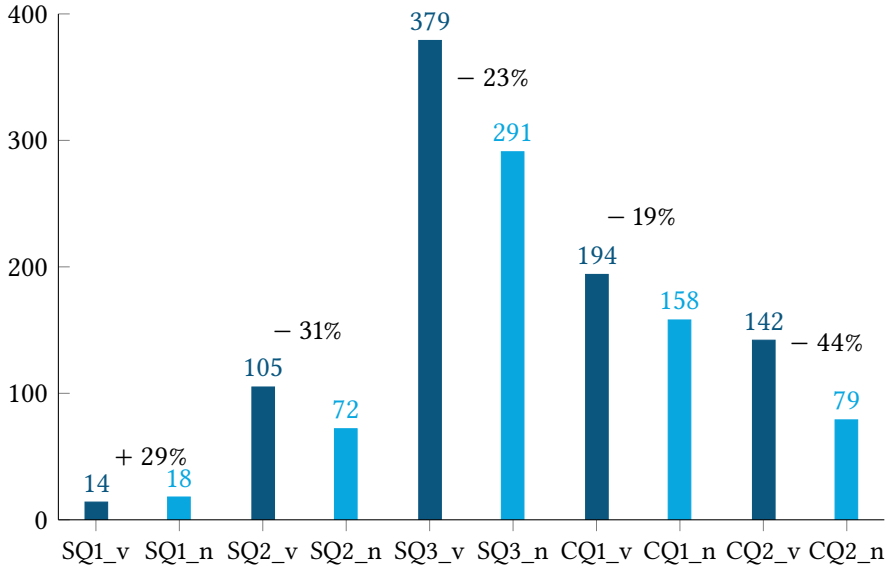
anzahl (im Vergleich zu der Gruppe Funktionsverb + Nominalphrase) verbunden. Ebenfalls zeigte die Humanevaluation, dass das Vermeiden der Funktionsverbgefüge im Falle der Gruppe Funktionsverb + Präpositionalphrase zu einer stärkeren Qualitätsverbesserung beitrug (+ 0,34 im Vergleich zu + 0,23 Qualitätsanstieg bei der Gruppe Funktionsverb + Nominalphrase) (Tabelle 5.36).[28]

Tabelle 5.36: Qualitätsdifferenz bei den untersuchten FVG-Formen

	FVG aus Funktionsverb + Präpositionalphrase	**FVG aus Funktionsverb + Nominalphrase**
Anzahl der Fälle	43 MÜ	36 MÜ
Durchschnittliche Diff. der Allg. Q (nach KS − vor KS)	+ 0,34	+ 0,23

Durch die Humanevaluation wird ferner ersichtlich (Abbildung 5.35), dass alle Qualitätskriterien sich nach KS verbesserten (insbesondere die Verständlichkeit (CQ2) sowie alle Stilqualitätskriterien).

[28] Die Qualitätsverbesserung bezieht sich hier auf die allgemeine Qualität, d. h. der Mittelwert der Stilqualität und der Inhaltsqualität.



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder.
CQ2: Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.35: „FVG verm." – Vergleich der Qualitätskriterien

In Tabelle 5.37 wurde der Kollokationsfehler in ‚does … be in the use' behoben, nachdem das Funktionsverbgefüge vermieden wurde (nach KS):

Daraufhin stiegen sowohl die Inhaltsqualität als auch die Stilqualität, wobei der Anstieg bei der Stilqualität höher war (SQdiff + 1,50 und CQdiff + 0,13 auf der Likert-Skala).

5.4.3.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung bei der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS − Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS − Q. vor KS) zusammenhängt. Mithilfe des Spearman-Tests konnten mehrere signifikante Korrelationen nachgewiesen werden:

Bei der Stilqualität gab es einen signifikanten negativen mittleren Zusammenhang zwischen der Differenz in dem lexikalischen Fehlertyp LX.4 „Zusätzliches

Tabelle 5.37: Beispiel 26

Vor-KS	**Steht** die Maschine nicht **im Einsatz**, den Hauptschalter auf "0" setzen.
RBMÜ Lucy	If the machine does not be in the use, set the main switch to "0".
Nach-KS	**Wird** die Maschine nicht **verwendet**, den Hauptschalter auf "0" setzen.
RBMÜ Lucy	If the machine is not used, set the main switch to "0".

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Wort eingefügt“ sowie dem semantischen Fehlertyp SM.11 „Verwechslung des Sinns“ und der Differenz in der Stilqualität; sowie einen signifikanten negativen schwachen Zusammenhang zwischen der Differenz im semantischen Fehlertyp SM.13 „Kollokationsfehler“ und der Differenz in der Stilqualität (Tabelle 5.38).

Bei der Inhaltsqualität gab es einen signifikanten negativen mittleren Zusammenhang zwischen der Differenz im semantischen Fehlertyp SM.12 „Falsche Wahl“ und der Differenz in der Inhaltsqualität; sowie einen signifikanten negativen schwachen Zusammenhang zwischen der Differenz in den Fehlertypen LX.3 (Lexik – Wort ausgelassen), LX.4 (Lexik – Zusätzliches Wort eingefügt) sowie SM.13 (Semantik – Kollokationsfehler) und der Differenz in der Inhaltsqualität (nächste Tabelle 5.38).

Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht erwiesen werden.

Diese signifikanten negativen Korrelationen deuten darauf hin, dass sobald die Fehleranzahl der genannten Fehlertypen sank, die Qualität stieg. Ein Beispiel hierfür ist der folgende Satz (Tabelle 5.39).

In diesem Beispiel kam bei der Verwendung des Funktionsverbgefüges der lexikalische Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ in ‚to make a selection‘ vor. Bei einer Formulierung des Satzes mit dem bedeutungstragenden Verb wurde der Fehler behoben. Daraufhin stiegen die Stil- und Inhaltsqualität jeweils um 1,00 Punkte auf der Likert-Skala.

Tabelle 5.38: „FVG verm." – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Differenz der Anzahl der **LX.3 „W. fehlt"**	84	,544	,067
Differenz der Anzahl der **LX.4 „W. extra"**	84	,005	−,306
Differenz der Anzahl der **SM.11 „Sinn"**	84	,003	−,317
Differenz der Anzahl der **SM.12 „f. Wahl"**	84	,184	−,146
Differenz der Anzahl der **SM.13 „Kollok."**	84	,014	−,267
Differenz CQ (nach KS − vor KS)			
Differenz der Anzahl der **LX.3 „W. fehlt"**	84	,008	−,286
Differenz der Anzahl der **LX.4 „W. extra"**	84	,023	−,248
Differenz der Anzahl der **SM.11 „Sinn"**	84	,111	−,175
Differenz der Anzahl der **SM.12 „f. Wahl"**	84	< ,001	−,375
Differenz der Anzahl der **SM.13 „Kollok."**	84	,048	−,216
Differenz allg. Q (nach KS − vor KS)			
Differenz der Anzahl der **LX.3 „W. fehlt"**	84	,023	−,248
Differenz der Anzahl der **LX.4 „W. extra"**	84	,004	−,307
Differenz der Anzahl der **SM.11 „Sinn"**	84	,016	−,262
Differenz der Anzahl der **SM.12 „f. Wahl"**	84	< ,001	−,385
Differenz der Anzahl der **SM.13 „Kollok."**	84	,027	−,242

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz	nicht signifikant (p ≥ 0,05)	ρ: Korrelationskoeffizient
schwache Korrelation (ρ >=0,1)	mittlere Korrelation (ρ >= 0,3)	starke Korrelation (ρ >= 0,5)

Tabelle 5.39: Beispiel 27

Vor-KS	Nach Ihrer Registrierung im Programm können Sie aus den Leistungen eine **Auswahl treffen**.
SMÜ SDL	After your registration in the program, you can select from the services to make a selection.
Nach-KS	Nach Ihrer Registrierung im Programm können Sie aus den Leistungen **wählen**.
SMÜ SDL	After your registration in the program, you can select from the services.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.3.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Auf MÜ-Systemebene stieg nach der Anwendung der KS-Regel nur die Stilqualität bei zwei Systemen signifikant (Tabelle 5.40): bei dem RBMÜ-System Lucy (+ 15,4 %) und dem HMÜ-System Systran (+ 13,8 %). Alle weiteren Veränderungen in der Stil- und Inhaltsqualität waren bei allen Systemen niedrig.

Wie die Aufteilung der Annotationsgruppen zeigte, waren 88 % der Übersetzungen des NMÜ-Systems Google Translate sowohl vor als auch nach der Anwendung der KS-Regel richtig (Annotationsgruppe RR). Entsprechend ist zu erwarten, dass das Qualitätsniveau sich nicht verändert. Ebenfalls waren die Ergebnisse bei dem HMÜ-System Bing sowie dem SMÜ-System SDL vor und nach der Anwendung der KS-Regel hinsichtlich der Annotationsgruppen gemischt, daher konnte keine signifikante Veränderung in der Qualität festgestellt werden (Tabelle 5.40).

5.4.3.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwies sich bei dem HMÜ-System Bing nur ein signifikanter starker negativer Zusammenhang zwischen dem lexikalischen Fehlertyp LX.3 „Wort ausgelassen“ und der Inhaltsqualität (Tabelle 5.41).

Bei dem RBMÜ-System Lucy erwies sich ein signifikanter starker negativer Zusammenhang zwischen dem semantischen Fehlertyp SM.11 „Verwechslung des

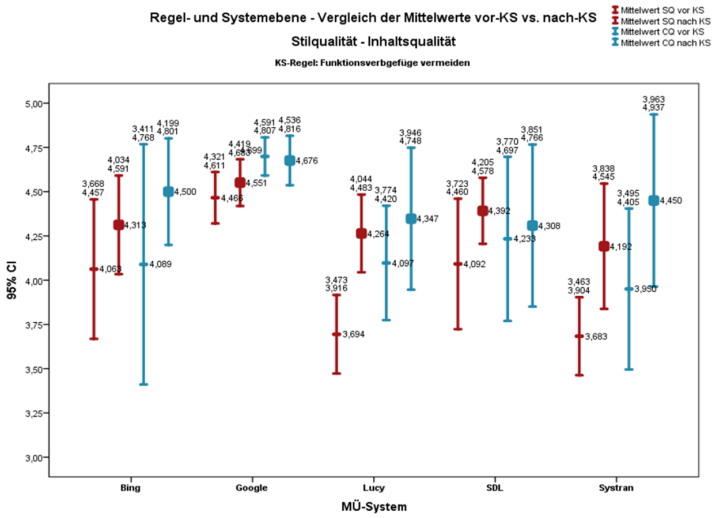


Abbildung 5.36: „FVG verm.“ – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.40: „FVG verm.“ – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	14	,431	− ,787	14	,362	−,912	14	,248	−1,155
Google	22	,392	−,857	22	,946	−,067	22	,536	−,618
Lucy	18	,002	− 3,030	18	,143	−1,463	18	,011	− 2,534
SDL	15	,055	−1,920	15	,850	−,189	15	,450	−,755
Systran	15	,035	− 2,108	15	,064	−1,855	15	,036	− 2,102

p: Signifikanz z: Teststatistik nicht signifikant ($p \geq 0{,}05$)

Sinns“ und der Stilqualität; sowie ein signifikanter starker negativer Zusammenhang zwischen dem semantischen Fehlertyp SM.12 „Falsche Wahl“ und der Inhaltsqualität (Tabelle 5.41).

Bei dem SMÜ-System SDL konnten mehrere signifikante Korrelationen nachgewiesen werden: Bei der Stilqualität gab es einen signifikanten negativen starken Zusammenhang zwischen der Differenz in den Fehlertypen LX.4 „Lexik – Zusätzliches Wort eingefügt“ (vgl. Tabelle 5.39) und GR.10 „Grammatik – Falsche Wortstellung“ einzeln und der Differenz in der Stilqualität (Tabelle 5.41), während es bei der Inhaltsqualität einen signifikanten negativen starken Zusammenhang zwischen der Differenz in den Fehlertypen LX.4 „Lexik – Zusätzliches Wort eingefügt“ und SM.12 „Semantik – Falsche Wahl“ einzeln und der Differenz in der Inhaltsqualität gab (Tabelle 5.41).

Bei dem HMÜ-System Systran erwies sich nur ein signifikanter starker negativer Zusammenhang zwischen dem grammatischen Fehlertyp GR.7 „Falsche Wortart / Wortklasse“ und der Inhaltsqualität (Tabelle 5.41).

Die Korrelation zwischen der Qualitätsdifferenz und der Fehleranzahldifferenz lässt sich anhand Tabelle 5.42 beleuchten: Der semantische Fehlertyp SM.12 in ‚suits in‘ wurde eliminiert, nachdem das bedeutungstragende Verb anstelle des Funktionsverbgefüges verwendet wurde. Daraufhin stiegen deutlich die Stilqualität um 1,38 Punkte und die Inhaltsqualität um 1,88 Punkte auf der Likert-Skala.

Die Bewerter fanden den semantischen Fehler (in ‚suits in‘) irreführend. Einer der Kommentare lautete: „The reader will have no idea what ‘suits in the responsibility of the planner’ means. The text therefore fails in its informative function. Using ‘suits’ can be confusing. I would therefore write: ‘It is the responsibility of the planner …’“

5.4.3.6 Vergleich der MÜ-Qualität mit vs. ohne Funktionsverbgefüge auf Annotationsgruppenebene

Mit Ausnahme der Gruppe FR fiel die Differenz in der Stil- und Inhaltsqualität[29] (nach KS – vor KS) bei allen anderen Annotationsgruppen gering aus.

In der Gruppe FF (Übersetzung vor und nach KS falsch) wurden die vor KS vorgekommenen Fehler – vor allem die semantischen Fehler – in vereinzelten Fällen nach der Anwendung der KS-Regel *zum Teil* eliminiert. Daher stieg die Qualität durchschnittlich nur leicht, demzufolge war die Differenz in der Qualität insignifikant (Tabelle 5.43).

In der Gruppe FR stiegen die Stil- und Inhaltsqualität erwartungsgemäß hochsignifikant (Tabelle 5.43): bei der Stilqualität ($z (N = 25) = -4{,}205 / p < {,}001$) bzw.

[29] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.41: „FVG verm.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Bing			Lucy			SDL			Systran		
	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ
Differenz der Anzahl SQ (nach KS − vor KS)												
LX.3 „W. fehlt“	14	,516	,190									
LX.4 „W. extra“							15	,020	−,594			
GR.7 „F. W.Art“										21	,060	− ,417
GR.10 „Wortst.“							15	,010	−,641			
SM.11 „Sinn“				18	,027	−,521						
SM.12 „F. Wahl“				18	,501	−,169	15	,078	−,468			
Differenz der Anzahl CQ (nach KS − vor KS)												
LX.3 „W. fehlt“	14	,035	−,566									
LX.4 „W. extra“							15	,009	−,646			
GR.7 „F. W.Art“										21	,018	−,511
GR.10 „Wortst.“							15	,295	,290			
SM.11 „Sinn“				18	,289	−,264						
SM.12 „F. Wahl“				18	,021	−,537	15	,020	−,593			
Differenz der Anzahl Q (nach KS − vor KS)												
LX.3 „W. fehlt“	14	,110	− ,446									
LX.4 „W. extra“							15	,007	−,664			
GR.7 „F. W.Art“										21	,018	−,509
GR.10 „Wortst.“							15	,081	−,465			
SM.11 „Sinn“				18	,053	−,463						
SM.12 „F. Wahl“				18	,014	−,567	15	,020	−,594			

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Tabelle 5.42: Beispiel 28

Vor-KS	Es **liegt in der Verantwortung** des Planers, aufeinander abgestimmte Produkte einzusetzen.
RBMÜ Lucy	It suits in the responsibility of the planner to use compatible products.
Nach-KS	Der Planer **ist dafür verantwortlich**, aufeinander abgestimmte Produkte einzusetzen.
RBMÜ Lucy	The planner is responsible for using compatible products.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

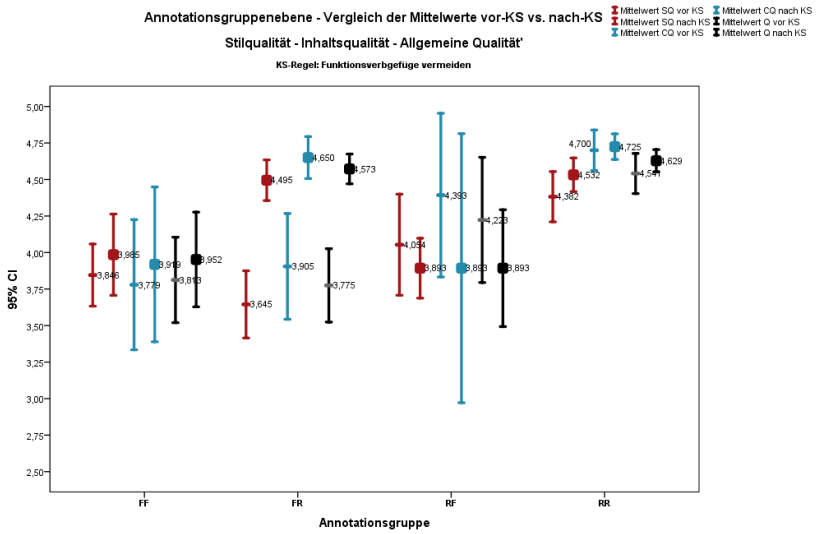


Abbildung 5.37: „FVG verm.“ – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

Tabelle 5.43: „FVG verm." – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	17	,420	−,807
Differenz CQ (nach KS − vor KS)	17	,636	−,474
Differenz allg. Q (nach KS − vor KS)	17	,421	−,805
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	25	< ,001	− 4,205
Differenz CQ (nach KS − vor KS)	25	< ,001	− 3,884
Differenz allg. Q (nach KS − vor KS)	25	< ,001	− 4,043
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	7	,416	,813
Differenz CQ (nach KS − vor KS)	7	,107	−1,612
Differenz allg. Q (nach KS − vor KS)	7	,058	−1,892
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	35	,140	−1,477
Differenz CQ (nach KS − vor KS)	35	,926	−,093
Differenz allg. Q (nach KS − vor KS)	35	,190	−1,312

bei der Inhaltsqualität (z (N = 25) = − 3,884/ p < ,001). Durch die Aufhebung der semantischen Fehler (meist SM.13 Kollokationsfehler), die mit der Verwendung des Funktionsverbgefüges verbunden sind, stiegen die Verständlichkeit und Genauigkeit sowie die Idiomatik und stilistische Adäquatheit[30] der MÜ deutlich. In Tabelle 5.44 war der Anstieg relativ markant und betrug bei der Stilqualität 1,38 Punkte bzw. bei der Inhaltsqualität 2,88 Punkte auf der Likert-Skala.

Tabelle 5.44: Beispiel 29

Vor-KS	Der Bediener darf erst die Maschine **in Betrieb nehmen**, wenn er die Betriebsanleitung gelesen hat.
HMÜ Systran	The operator may only take the machine in enterprise after reading the operating instructions.
Nach-KS	Der Bediener darf erst die Maschine **starten**, wenn er die Betriebsanleitung gelesen hat.
HMÜ Systran	The operator may only start the machine after reading the operating instructions.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

In einem weiteren Satz (Tabelle 5.45) konnte das MÜ-System (vor KS) das Kompositum ‚Fleckenbehandlung', aus dem das Funktionsverbgefüge zum Teil besteht, richtig übersetzen. Allerdings beinhaltete die Übersetzung einen Kollokationsfehler (in ‚treatment …accomplished'), der nach der Verwendung des bedeutungstragenden Verbs ‚behandeln' (nach KS) behoben werden konnte.

In der Gruppe RF war die Anzahl der vorgekommenen Fälle relativ klein (8 %). In dieser Gruppe wurde das Funktionsverbgefüge (vor KS) richtig übersetzt, während das bedeutungstragende Verb aus unterschiedlichen Gründen falsch übersetzt wurde. Folgender Satz ist ein Beispiel für solche Fälle, bei denen die Stilqualität um 0,25 Punkte bzw. die Inhaltsqualität um 0,63 Punkte auf der Likert-Skala zurückgingen (Tabelle 5.46).

In der Gruppe RR (Übersetzung vor und nach KS richtig) war die Inhaltsqualität vor und nach der Anwendung der KS-Regel vergleichbar. Auf der anderen Seite stieg die Stilqualität in manchen Fällen, da die Übersetzung des bedeutungstragenden Verbs von den Bewertern als prägnanter wahrgenommen wurde. In

[30] Stilistische Adäquatheit im Sinne von Hutchins & Somers (1992: 163) „the extent to which the translation uses the language appropriate to its content and intention". Mehr zu den Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.45: Beispiel 30

Vor-KS	Die Fleck**behandlung** muss so schnell wie möglich **durchgeführt** werden.
HMÜ Systran	The stain treatment should be accomplished as soon as possible.
Nach-KS	Die Flecke**n** müssen so schnell wie möglich **behandelt** werden.
HMÜ Systran	The stains should be treated as soon as possible.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Tabelle 5.46: Beispiel 31

Vor-KS	**Die Abwicklung von Garantieleistungen erfolgt über die lokale Service-Hotline.**
SMÜ SDL	Handling of warranty services is effected by the local service hotline.
Nach-KS	**Die Garantieleistungen werden über die lokale Service-Hotline abgewickelt.**
SMÜ SDL	The warranty services are XXX by the local service hotline.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Tabelle 5.47 blieb die Inhaltsqualität unverändert, während die Stilqualität um 1,13 Punkte auf der Likert-Skala stieg.

5.4.3.7 Vergleich der AEM-Scores mit vs. ohne Funktionsverbgefüge sowie die Korrelation zwischen den AEM-Scores und der Qualität

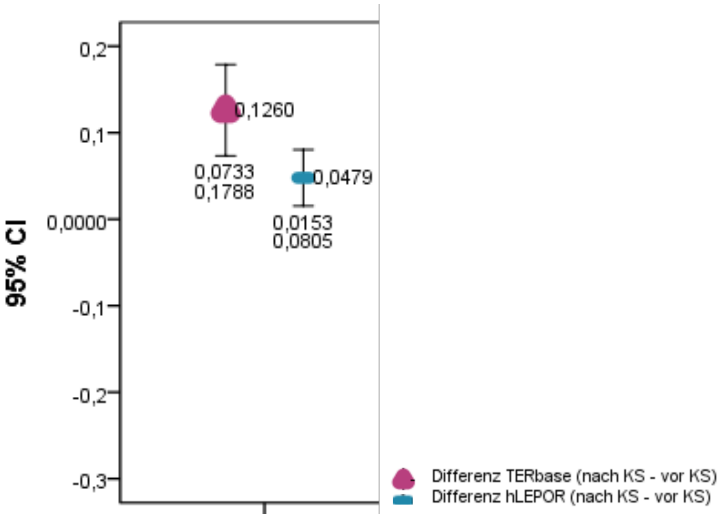
Der Vergleich der AEM-Scores mit und ohne Funktionsverbgefüge zeigte sowohl mit TERbase als auch mit hLEPOR eine Verbesserung der AEM-Scores.

Der Mittelwert der Differenz (nach KS – vor KS) im AEM-Score pro Satz lag für TERbase bei ,126 (SD = ,244) und für die hLEPOR bei ,048 (SD = ,151) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben). Die Differenzen

Tabelle 5.47: Beispiel 32

Vor-KS	Die Höhen**verstellung** der Fronten können Sie mittels eines Schraubendrehers **vornehmen**.
HMÜ Systran	You can make the height adjustment of the fronts using a screwdriver.
Nach-KS	Die **Höhe** der Fronten können Sie mittels eines Schraubendrehers **verstellen**.
HMÜ Systran	You can adjust the height of the fronts using a screwdriver.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.38: „FVG verm.“ – Mittelwert der Differenz der AEM-Scores

(nach KS − vor KS) in TERbase und hLEPOR erwiesen sich als signifikant (z (N = 85) = − 4,353 / p < ,001) bzw. (z (N = 85) = − 2,975 / p = ,003). Dieses Ergebnis weist darauf hin, dass – nachdem das Funktionsverbgefüge vermieden wurde (nach KS) –weniger Edits erforderlich waren.

5.4.3.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter starker positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität. Tabelle 5.48 präsentiert die Korrelationswerte:

Tabelle 5.48: „FVG verm." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	84	< ,001	,503	starker Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	84	< ,001	,510	starker Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Dieses Ergebnis deutet darauf hin, dass – nachdem das bedeutungstragende Verb anstelle des Funktionsverbgefüges verwendet wurde – die Scores der beiden AEMs sich verbesserten und die Qualität stieg.

5.4.3.8 Analyse der zweiten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der For-

schungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der zweiten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (–) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[31]

Regel 2: Funktionsverbgefüge vermeiden

Erster Analysefaktor: Vergleich der Fehleranzahl mit vs. ohne Funktionsverbgefüge

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene: H0 wurde abgelehnt und somit H1 bestätigt. Die Fehleranzahl sank signifikant, nachdem das Funktionsverbgefüge vermieden wurde.	**Anz.F. (–)**
Auf Regel- und MÜ-Systemebene: Bei Lucy, SDL und Systran sank die Fehleranzahl signifikant, nachdem das Funktionsverbgefüge vermieden wurde.	**Lu (–)** **SD (–)** **Sy (–)**
Bei Bing und Google sank ebenfalls die Fehleranzahl, jedoch war der Rückgang nicht signifikant.	**Bi (–)** **Go (–)**

[31] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Zweiter Analysefaktor

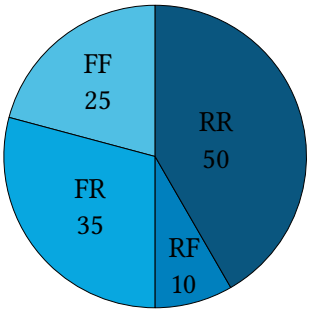


Abbildung 5.39: Aufteilung der Annotationsgruppen auf Regelebene

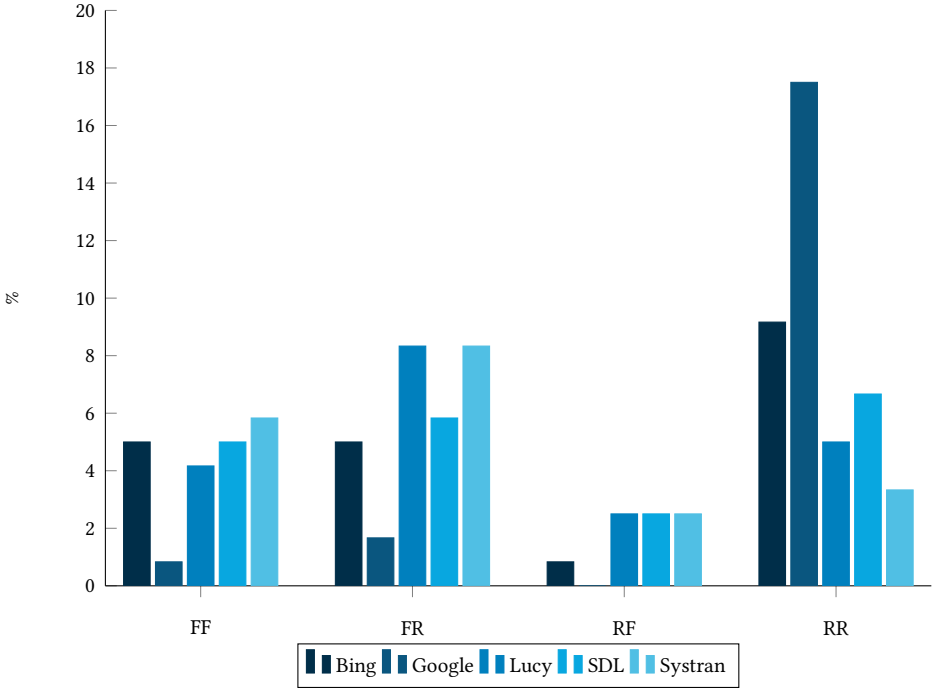


Abbildung 5.40: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Dritter Analysefaktor: Vergleich der Fehlertypen mit vs. ohne Funktionsverbgefüge

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H1 wurde nur für einen Fehlertyp bestätigt.
Die Fehleranzahl von SM.13 „Kollokationsfehler" sank signifikant, nachdem das Funktionsverbgefüge vermieden wurde.

SM.13 (–)

Auf Regel- und MÜ-Systemebene:
Bei Lucy und Systran sank die Fehleranzahl von SM.13 „Kollokationsfehler" signifikant, nachdem das Funktionsverbgefüge vermieden wurde.

SM.13 (–):
Lu
Sy

Alle weiteren Veränderungen waren nicht signifikant.

Vierter Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne Funktionsverbgefüge

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Sowohl die Stil- als auch die Inhaltsqualität stiegen nach KS signifikant.

SQ (+)
CQ (+)

Auf Regel- und MÜ-Systemebene:
Nur die Stilqualität stieg bei Lucy und Systran signifikant.

SQ (+)
Lu Sy

Die Stilqualität stieg bei allen anderen Systemen, allerdings nicht signifikant.
Die Inhaltsqualität stieg bei allen Systemen – mit Ausnahme von Google Translate – nicht signifikant. Bei Google Translate sank sie leicht.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde nur für drei Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und des SM.11 „Verwechslung des Sinns“ einzeln und der Differenz der Stilqualität sowie ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des SM.12 „Falsche Wahl“ und der Differenz der Inhaltsqualität.

***neg* LX.4 <> SQ**
***neg* SM.11 <> SQ**
***neg* SM.12 <> CQ**

Auf Regel- und MÜ-Systemebene:
Bei Bing bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“ und der Differenz der Inhaltsqualität.

Bi
***neg* LX.3 <<>> CQ**

Bei Lucy bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des SM.11 „Verwechslung des Sinns“ und der Differenz der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des Fehlertyps 12 „SM – Falsche Wahl“ und der Differenz der Inhaltsqualität.

Lu
***neg* SM.11 <<>> SQ**
***neg* SM.12 <<>> CQ**

Bei SDL bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und des GR.10 „Falsche Wortstellung“ einzeln und der Differenz der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und des SM.12 „Falsche Wahl“ einzeln und der Differenz der Inhaltsqualität.

SD
***neg* LX.4 <<>> SQ**
***neg* GR.10 <<>> SQ**
***neg* LX.4 <<>> CQ**
***neg* SM.12 <<>> CQ**

Bei Systran bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.7 „Falsche Wortart / Wortklasse“ und der Differenz der Inhaltsqualität.

Sy
***neg* GR.7 <<>> CQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne Funktionsverbgefüge auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde nur für die Gruppe FR bestätigt:
Bei der Annotationsgruppe FR stiegen die Stil- und Inhaltsqualität signifikant, nachdem das Funktionsverbgefüge vermieden wurde.

SQ (+)
CQ (+)

Bei den Annotationsgruppen FF und RR stiegen die Stil- und Inhaltsqualität leicht.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe RF sanken die Stil- und Inhaltsqualität leicht.	**SQ (−)** **CQ (−)**

Siebter Analysefaktor: Vergleich der AEM-Scores mit vs. ohne Funktionsverbgefüge

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt. AEM-Scores sowohl von TERbase als auch von hLEPOR verbesserten sich signifikant, nachdem das Funktionsverbgefüge vermieden wurde.	**TERbase (+)** **hLEPOR (+)**

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS − vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter starker positiver Zusammenhang zwischen den Differenzen der Scores der beiden AEMs (TERbase und hLEPOR) und der Differenz der allgemeinen Qualität.

pos **TERbase <<>> Q**
pos **hLEPOR <<>> Q**

5.4.4 DRITTE REGEL: Konditionalsätze mit ‚Wenn' einleiten

5.4.4.1 Überblick

Im Folgenden wird die KS-Regel „Konditionalsätze mit ‚Wenn' einleiten" (auch bekannt als „Bedingungen als ‚Wenn'-Sätze formulieren") kurz beschrieben.[32] Zudem wird zusammenfassend und anhand von Beispielen demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Konditionalsätze mit ‚Wenn' einleiten (tekom-Regel-Nr. S 201 „Bedingungen als ‚Wenn'-Sätze formulieren")

Nach dieser Regel (tekom 2013: 66) sollen Bedingungssätze mit der Konjunktion „Wenn" oder „Falls" eingeleitet werden.

Begründung: Durch die Satzstruktur „Wenn-Nebensatz-Hauptsatz" wird das „Bedingung-Folge-Verhältnis" ersichtlich und somit die Textverständlichkeit erhöht (ebd.: 67).

Umsetzungsmuster:

Vor KS: Der Satz beginnt mit dem Verb.
Nach KS: Der Satz ist mit ‚Wenn' am Satzanfang formuliert.
Wenn der Nebensatz vor KS mit ‚so' formuliert ist, wurde ‚so' aus stilistischen Gründen nach KS entfernt (siehe Beispiel unten).

KS-Stelle

Vor KS: das Verb
Nach KS: ‚Wenn' + das Verb

[32] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

Beispiele

Ist die Seriennummer des Gerätes bekannt, kann im Feld ...

Wenn die Seriennummer des Gerätes bekannt ist, kann im Feld ...

Werden die vordefinierten Werte verändert, so erfolgt die Umrechnung automatisch.

Wenn die vordefinierten Werte verändert werden, erfolgt die Umrechnung automatisch.

Aufteilung der Testsätze: Da die Konditionalsätze im Deutschen mit unterschiedlichen Verben beginnen können und dies wiederum unterschiedliche Schwierigkeitsgrade für die MÜ-Systeme bedeutet, bestehen die Testsätze aus:

7 Konditionalsätzen, die mit dem Verb ‚Werden' beginnen,

8 Konditionalsätzen mit dem Verb ‚Sind' und

9 Konditionalsätzen mit weiteren unterschiedlichen Verben.

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.4.2 Vergleich der Fehleranzahl bei Konditionalsätzen mit vs. ohne ‚Wenn'

Die Fehleranzahl sank deutlich um 56,5 % von 92 Fehlern im Falle der Formulierung der Bedingungen mit Verben (M = ,77 / SD = 1,059 / N = 120) auf 40 Fehler bei der Formulierung mit ‚Wenn' (M = ,33 / SD = ,626 / N = 120), siehe Abbildung 5.41 und Abbildung 5.42. Der Mittelwert der Differenz (nach KS − vor KS) der Fehleranzahl pro Satz lag somit bei − ,43 (SD = 1,019) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,62 (SD =,814) und einem Maximum von − ,25 (SD = 1,197) (Bootstrapping mit 1000 Stichproben). Die Differenz (nach KS − vor KS) der Fehleranzahl erwies sich als hochsignifikant (z (N = 120) = − 4,282 / p < ,001).

Im Datensatz ist ersichtlich, dass das Einleiten eines Konditionalsatzes mit einem Verb den MÜ-Systemen Probleme bereitet. Knapp 80 % (19 von 24) der analysierten Sätze wurden von mindestens einem MÜ-System falsch übersetzt und mithilfe der KS-Regel korrigiert. Daher war die Anwendung der KS-Regel im Hinblick auf die Fehleranzahl sinnvoll. Wie in der Überblickstabelle erwähnt, bestehen die Testsätze aus sieben Konditionalsätzen, die mit dem Verb ‚Werden' beginnen, acht Konditionalsätzen mit dem Verb ‚Sind' und neun Konditionalsätzen mit weiteren unterschiedlichen Verben am Satzanfang. Die Ergebnisse bei

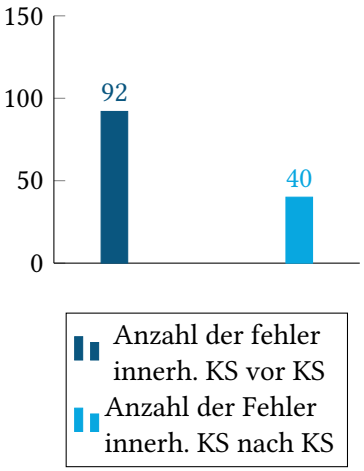


Abbildung 5.41: „Kondi. m. Wenn“ – Fehlersumme vor vs. nach KS

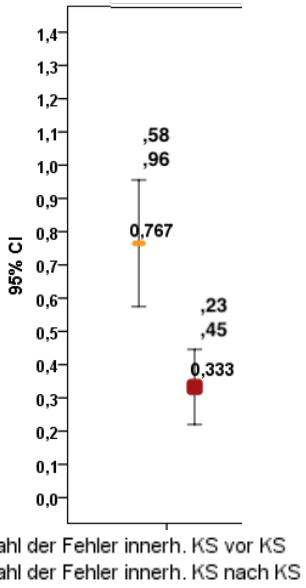


Abbildung 5.42: „Kondi. m. Wenn“ – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

den drei Gruppen waren auf Regelebene im Allgemeinen relativ ähnlich (Tabelle 5.49). Nur bei der Kategorie *Fehler vollständig korrigiert* sind mehr korrigierte Fälle bei den Konditionalsätzen mit ‚Sind‘ im Vergleich zu den anderen Verben zu beobachten:

Tabelle 5.49: Fehleranzahlveränderung bei den verschiedenen Konditionalsätzen

Veränderung in der Fehleranzahl nach Anwendung der Regel	Konditionalsätze mit ‚Sind‘	Konditionalsätze mit untersch. Verben	Konditionalsätze mit ‚Werden‘
Fehler vollständig korrigiert	16 MÜ	12 MÜ	10 MÜ
Fehleranzahl gleich geblieben	20 MÜ	29 MÜ	22 MÜ
Fehleranzahl stieg	4 MÜ	4 MÜ	3 MÜ

N = 120

Die verschiedenen Annotationsgruppen werden unter §5.4.4.3 näher betrachtet.

5.4.4.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Die Fehleranzahl nach der Umsetzung der KS-Regel sank bei allen Systemen mit Ausnahme des NMÜ-Systems: Google Translate war in der Lage 22 der 24 analysierten Sätze sowohl vor als auch nach der Umsetzung der KS-Regel korrekt zu übersetzen. Bei den zwei falschen Übersetzungen (Abbildung 5.43) handelt es sich um einen semantischen Fehler (‚Wenn' wurde als ‚When' anstatt ‚If' übersetzt) (Tabelle 5.50):

Tabelle 5.50: Beispiel 33

Vor-KS	**Ist** diese Zeit **erreicht**, muss das Gerät für 2 Minuten abkühlen.
GNMÜ	When this time is reached, the unit must cool down for 2 minutes.
Nach-KS	**Wenn** diese Zeit **erreicht ist**, muss das Gerät für 2 Minuten abkühlen.
GNMÜ	When this time is reached, the unit must cool down for 2 minutes.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Die beiden falsch übersetzten Sätze waren sehr kontextabhängig, da das Konditionalverb ohne Kontext sowohl als ‚If' als auch ‚When' übersetzt werden kann. Mehr zu den Fehlertypen ist unter §5.4.4.4 aufgeführt.

Weitere nicht signifikante Differenzen in der Fehleranzahl waren bei dem RBMÜ-System Lucy (Mdiff = − ,125) und dem hybriden System Systran (Mdiff = − ,042) zu finden (Abbildung 5.43).

Hingegen war die Differenz bei dem HMÜ-System Bing (Mdiff = − 1,208; z (N = 24) = − 3,859 / p < ,001) und dem SMÜ-System SDL (Mdiff = − ,833; z (N = 24) = − 2,357 / p = ,018) signifikant (Abbildung 5.43). Bei diesen beiden Systemen war die Anwendung der KS-Regel sehr nützlich. Die korrigierten Fehlertypen werden im §5.4.4.4 detaillierter dargestellt.

5.4.4.3 Aufteilung der Annotationsgruppen

Genau die Hälfte der übersetzten Sätze war sowohl bei der Formulierung des Konditionalsatzes mit Verben (vor KS) als auch mit ‚Wenn' (nach KS) fehlerfrei

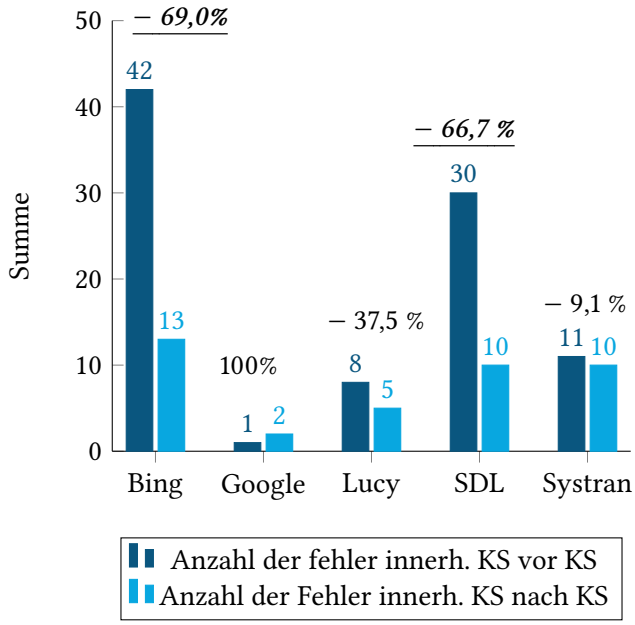


Abbildung 5.43: „Kondi. m. Wenn“ – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

(Gruppe RR) (Abbildung 5.46). Knapp ein Viertel der Sätze beinhaltete vor der Anwendung der KS-Regel Fehler, die nach der Anwendung der Regel vollständig korrigiert wurden (Gruppe FR) (Abbildung 5.47). Die drittgrößte Gruppe war die Gruppe FF. Diese Gruppe repräsentiert fast 21 % der analysierten Sätze (Abbildung 5.48). Hier beinhaltete die Übersetzung Fehler sowohl vor der Anwendung der KS-Regel als auch danach.

Schließlich wurden nur 5 % der Sätze bei der Formulierung des Konditionalsatzes mit Verb fehlerfrei übersetzt und erst nach der Formulierung mit ‚Wenn‘ falsch übersetzt (Abbildung 5.44). Im §5.4.4.4 werden die Fehlertypen näher betrachtet.

5.4.4.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Bei der dominanten Annotationsgruppe RR verzeichnete das NMÜ-System Google Translate den höchsten Prozentsatz mit 92 %, gefolgt von dem RBMÜ-System

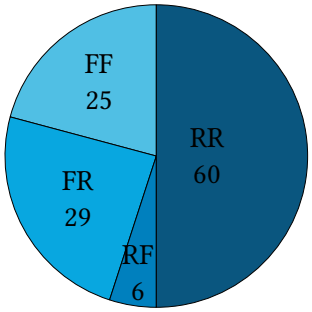


Abbildung 5.44: „Kondi. m. Wenn“ – Aufteilung der Annotationsgruppen

Lucy mit 71 % und schließlich von dem HMÜ-System Systran mit 58 % (Abbildung 5.45).

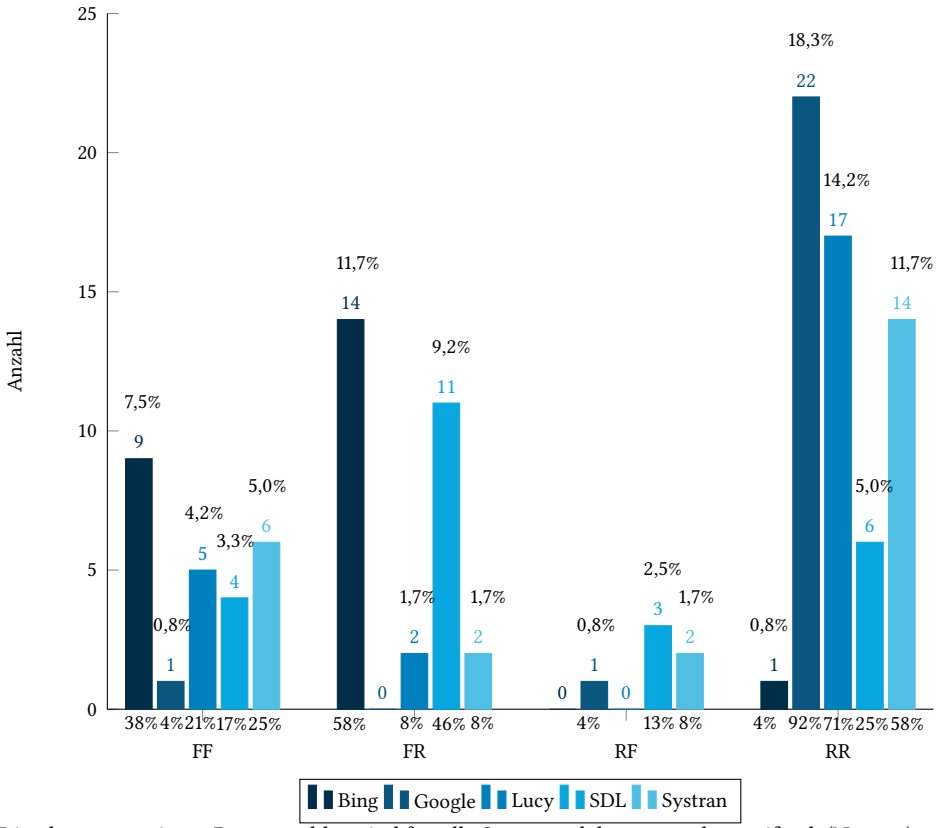
Erneut erzielten die HMÜ-Systeme unterschiedliche Ergebnisse. Anders als Systran waren 58 % der Übersetzungen bei dem HMÜ-System Bing vor der Anwendung der KS-Regel falsch und danach richtig (Annotationsgruppe FR) (Abbildung 5.45). Eine Formulierung der Bedingungen mit ‚Wenn‘ hat das System entsprechend unterstützt. Auch das SMÜ-System SDL erzielte ein ähnliches Ergebnis mit 46 % aus der Annotationsgruppe FR (Abbildung 5.45).

5.4.4.4 Vergleich der Fehlertypen bei Konditionalsätzen mit vs. ohne ‚Wenn‘

Das Einleiten eines Konditionalsatzes mit einem Verb (vor KS) wurde in vielen Fällen von den MÜ-Systemen fehlerhaft übersetzt. Die Systeme übersetzen dieses Verb in den meisten Fällen falsch bzw. doppelt. Somit entstehen zwei lexikalische Fehlertypen: Fehlertyp LX.3 durch das Auslassen der Konjunktion ‚Wenn‘ und Fehlertyp LX.4 durch das Hinzufügen eines überflüssigen Verbs (‚is‘ in Tabelle 5.51):

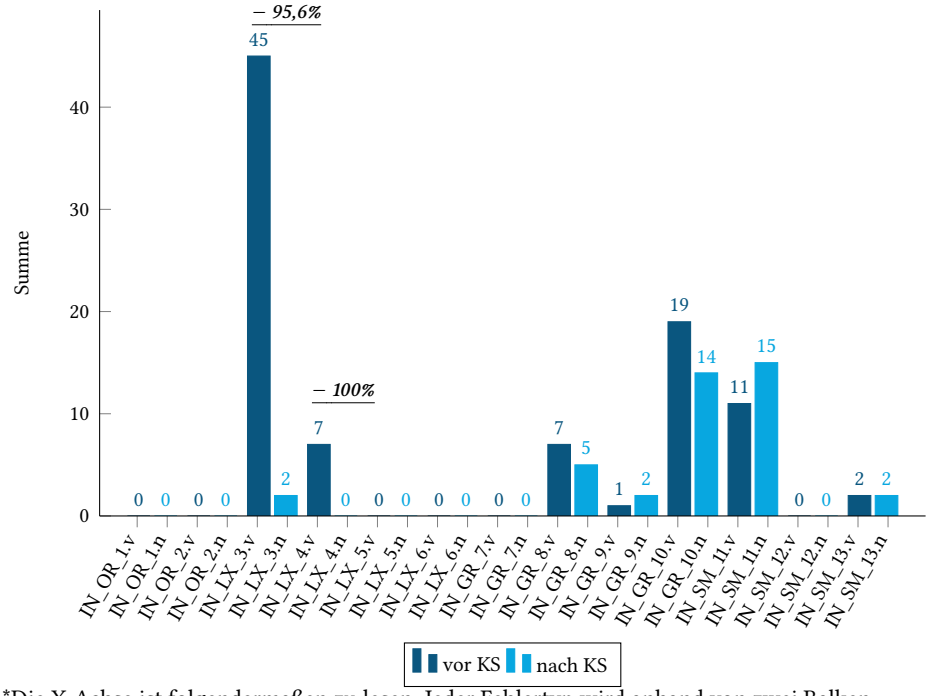
Die Fehleranzahl beider lexikalischen Fehlertypen LX.3 „Wort ausgelassen“ und LX.4 „Zusätzliches Wort eingefügt“ sank nach der Formulierung der Konditionalsätze mit ‚Wenn‘: Bei dem Fehlertyp LX.3 sank die Fehleranzahl massiv von 45 auf 2 (– 95,6 % / Mv = ,38 / SDv = ,486 / Mn = ,02 / SDn = ,129 / N = 120) und bei dem Fehlertyp LX.4 wurden alle 7 vorgekommenen Fehler vor KS nach KS vollständig behoben (– 100 % / Mv = ,06 / SDv = ,235 / Mn = – / SDn = – / N = 120) (Abbildung 5.46). Entsprechend erwies sich der Unterschied bei den beiden Fehlertypen als hochsignifikant ($p < ,001$ bzw. $p = ,023$ / N = 120).

Außerdem stellt die Übersetzung der Konditionalkonjunktion als Verb (‚Is‘ in Tabelle 5.51) am Satzbeginn einen semantischen Fehler dar, nämlich Fehlertyp



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.45: „Kondi. m. Wenn“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler vor KS und der zweite die Summe der Fehler nach KS, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
__*Signifikante Differenz vor vs. nach KS***__
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.46: „Kondi. m. Wenn“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

Tabelle 5.51: Beispiel 34

Vor-KS	**Ist** die Seriennummer des Gerätes bekannt, kann im Feld Seriennummer diese Nummer eingegeben werden.
SMÜ SDL	XXX Is the serial number of the device is known, this number can be entered in the "Serial Number" field.
Nach-KS	**Wenn** die Seriennummer des Gerätes bekannt **ist**, kann im Feld Seriennummer diese Nummer eingegeben werden.
SMÜ SDL	If the serial number of the device is known, this number can be entered in the "Serial Number" field.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

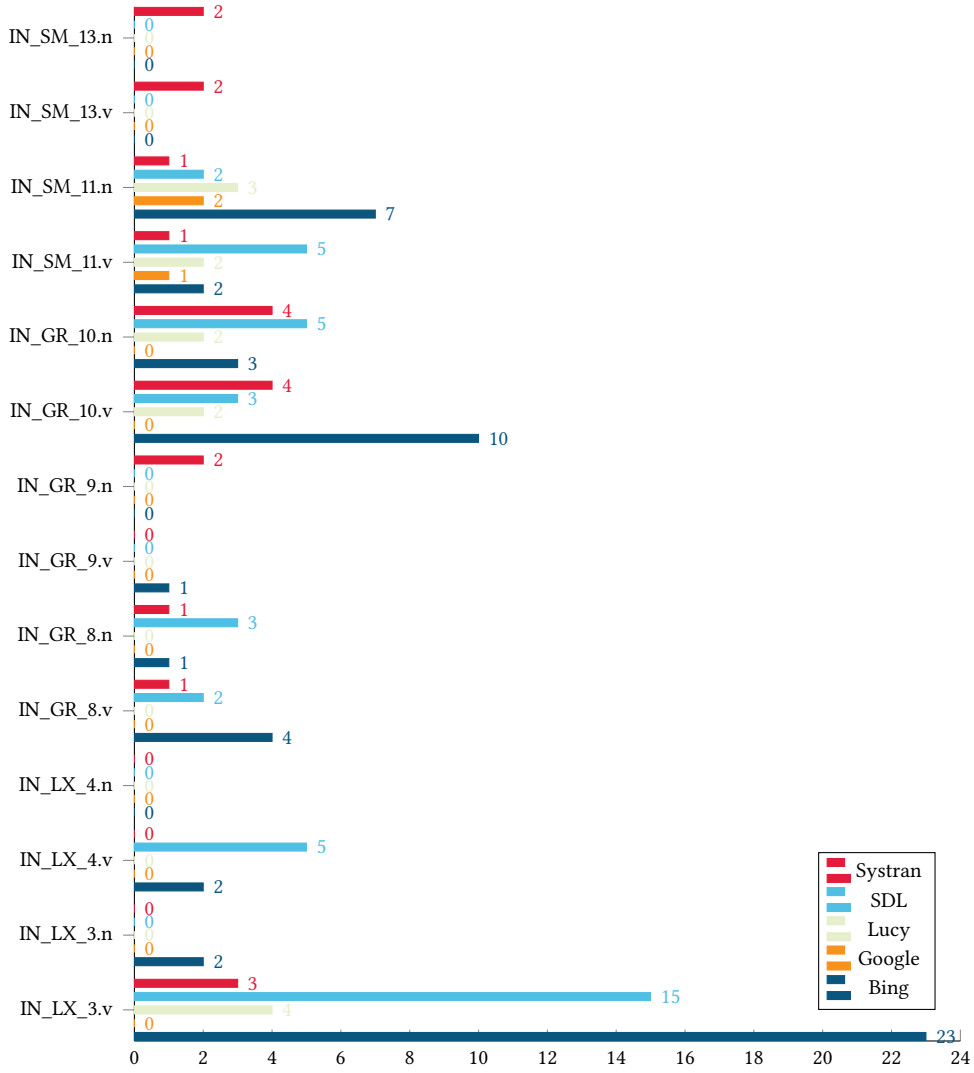
SM.11 „Verwechslung des Sinns“. Dieser Fehlertyp stieg systemübergreifend minimal (11 vor der Umsetzung bzw. 15 nach der Umsetzung der KS-Regel) und entsprechend war die Veränderung insignifikant. Der Grund für den Anstieg der Fehleranzahl liegt darin, dass ‚Wenn‘ in einigen Fällen als ‚When‘ anstatt ‚If‘ übersetzt wurde. Im kommenden Abschnitt wird dieser Punkt systembezogen anhand eines Beispiels näher erläutert.

5.4.4.5 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Während auf Regelebene die Fehleranzahl des Fehlertyps LX.3 „Lexik – Wort ausgelassen“ und LX.4 „Lexik – Zusätzliches Wort eingefügt“ sich signifikant veränderte, zeigt eine nähere Analyse der Fehlertypen bei den verschiedenen MÜ-Systemen Folgendes (Abbildung 5.47): Die Fehlertypen LX.3 „Lexik – Wort ausgelassen“ und GR.10 „Grammatik – Falsche Wortstellung“ wiesen einzeln bei dem HMÜ-System Bing signifikante Veränderungen auf. Bei dem SMÜ-System SDL veränderte sich nur Fehlertyp LX.3 „Lexik – Wort ausgelassen“ signifikant. Fehlertyp LX.4 „Lexik – Zusätzliches Wort eingefügt“ zeigte bei keinem bestimmten System eine signifikante Veränderung.

Fehlertyp LX.3 sank bei Bing von 23 auf 2 Fehler (– 91,3 %) und bei SDL von 15 auf 5 Fehler (– 66,7 %) (Abbildung 5.47). Fehlertyp GR.10 sank bei Bing von 10 auf 3 Fehler (– 70 %) (Abbildung 5.47). Entsprechend erwies sich der Unterschied in der Fehleranzahl der beiden Fehlertypen LX.3 und GR.10 wie folgt als signifikant (Tabelle 5.52).

Diese signifikante Veränderung ist in Tabelle 5.53 zu beobachten.



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 1, 2, 5, 6, 7 und 12 bei gar keinem MÜ-System vor.

Abbildung 5.47: „Kondi. m. Wenn“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.52: „Kondi. m. Wenn“ – Fehlertypen mit signifikanter Veränderung nach KS

	N	Mittelwert	Standard-abweichung	Signifikanz (McNemar-Test)
LX.3 „Wort ausgelassen“				
Bing	24 x 2	vor KS = ,96 nach KS = ,08	vor KS = ,204 nach KS = ,282	p < ,001
SDL	24 x 2	vor KS = ,63 nach KS = –	vor KS = ,495 nach KS = –	p < ,001
GR.10 „Falsche Wortstellung“				
Bing	24 x 2	vor KS = ,42 nach KS = ,13	vor KS = ,504 nach KS = ,338	p = ,016

Tabelle 5.53: Beispiel 35

Vor-KS	**Steht** die Maschine nicht im Einsatz, den Hauptschalter auf "0" setzen.
HMÜ Bing	XXX Is the machine not in use, set the main switch to "0".
Nach-KS	**Wenn** die Maschine nicht im Einsatz **steht**, den Hauptschalter auf "0" setzen.
HMÜ Bing	If the machine is not in use, ***set*** the main switch to "0".

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Hierbei wurden der lexikalische Fehlertyp LX.3 (in dem Auslassen von ‚If‘) und der grammatische Fehlertyp GR.10 (in der falschen Wortstellung des Verbs ‚Is‘) nach der Anwendung der KS-Regel behoben.

Der insignifikante Anstieg in der Fehleranzahl bei Fehlertyp SM.11 „Semantik – Verwechslung des Sinns“ bei dem HMÜ-System Bing und dem NMÜ-System Google kann durch Tabelle 5.54 verdeutlicht werden.

Solche semantischen Fehler der Verwechslung des Sinns wurden auch in den Ergebnissen der Humanevaluation beobachtet. Hierbei kommentierten die Teilnehmer, dass sie Kontextinformationen benötigen, um den korrekten Sinn erkennen zu können.

Tabelle 5.54: Beispiel 36

Vor-KS	**Schließt** der Kontaktschalter, so wird der Raumdruck-Sollwert aktiv.
GNMÜ	If the contact switch closes, the room pressure setpoint becomes active.
Nach-KS	**Wenn** der Kontaktschalter **schließt**, wird der Raumdruck-Sollwert aktiv.
GNMÜ	When the contact switch closes, the room pressure setpoint becomes active.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.4.6 Vergleich der MÜ-Qualität bei Konditionalsätzen mit vs. ohne ‚Wenn' sowie die Korrelation zwischen den Fehlertypen und der Qualität

Sowohl die Stil- als auch die Inhaltsqualität[33] stiegen nach der Formulierung der Konditionalsätze mit ‚Wenn' (nach KS). Auf den ersten Blick erkennt man in Abbildung 5.49, dass der Einfluss auf die Inhaltsqualität im Vergleich zur Stilqualität größer war. Angesichts des deutlichen Rückgangs der lexikalischen Fehlertypen LX.3 „Wort ausgelassen" und LX.4 „Zusätzliches Wort eingefügt" nach der Verwendung der Konjunktion ‚Wenn' (nach KS) (siehe §5.4.4.4), ist es nachvollziehbar, dass die Inhaltsqualität durch die erhöhte Verständlichkeit und Genauigkeit primär beeinflusst wird:

Die Stilqualität stieg um 3,8 % (Mv = 4,22 / SDv = ,533 / Mn = 4,38 / SDn = ,483 / N = 84). Die Inhaltsqualität stieg um 9,3 % (Mv = 4,21 / SDv = ,748 / Mn = 4,60 / SDn = ,599 / N = 84) (Abbildung 5.48). Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei ,168 (SD = ,562) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,046 und einem Maximum von ,290 und für die Inhaltsqualität bei ,384 (SD = ,800) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,210 und einem Maximum von ,558 (Bootstrapping mit 1000 Stichproben) (Abbildung 5.49). Die Differenzen (nach KS − vor KS) in der Stil- und Inhaltsqualität erwiesen sich als signifikant (z (N = 84) = − 2,585 / p = ,010) bzw. (z (N = 84) = − 3,983 / p < ,001).

[33] Definitionen der Qualität unter §4.5.5.1.

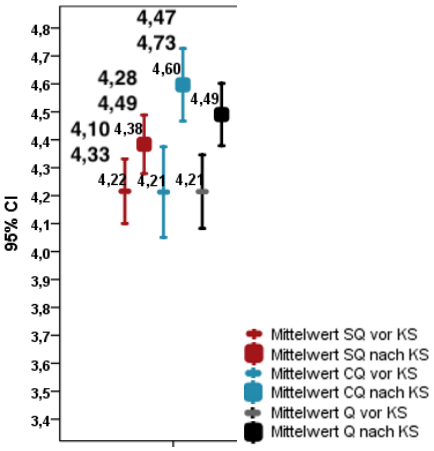


Abbildung 5.48: „Kondi. m. Wenn“ – Mittelwerte der Qualität vor und nach KS

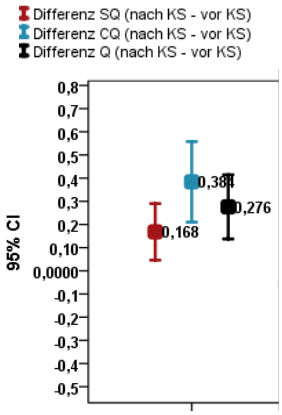


Abbildung 5.49: „Kondi. m. Wenn“ – Mittelwert der Qualitätsdifferenzen

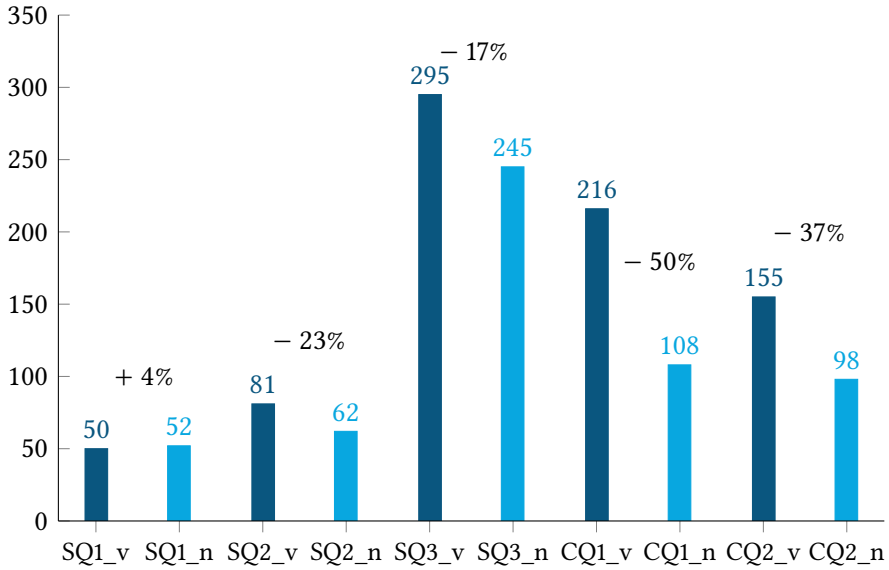
Durch die Humanevaluation wird ersichtlich, dass der Anstieg der allgemeinen Qualität durch den Anstieg der Inhaltsqualität begründet ist. Abbildung 5.50 „Vergleich der Qualitätskriterien“ zeigt, dass sich beide Inhaltsqualitätskriterien (CQ1 und CQ2) nach der Verwendung der Konjunktion ‚Wenn‘ anstelle von Verben am Satzbeginn deutlich verbesserten.

Wie Tabelle 5.55 zeigt, ist die MÜ vor KS unverständlich (CQ2) und gibt den Inhalt des Ausgangstexts nicht wieder (CQ1).

Tabelle 5.55: Beispiel 37

Vor-KS	**Ist** nur ein Gerät **angeschlossen**, so ist die Funktion PP zu wählen.
HMÜ Systran	Only if one device is attached, the "PP" function must be selected.
Nach-KS	**Wenn** nur ein Gerät **angeschlossen ist**, ist die Funktion PP zu wählen.
HMÜ Systran	If only one device is attached, the "PP" function must be selected.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder.
CQ2: Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.50: „Kondi. m. Wenn“ – Vergleich der Qualitätskriterien

Durch den Wortstellungsfehler in ‚If‘ (vor KS) wird falscher Inhalt vermittelt. Dieser Einfluss ist in den Kommentaren der Bewerter wiederzufinden „only ‚if‘ is wrong and actually provides false information, I suggest ‚if only‘“ (Kommentar eines Bewerters).

5.4.4.6.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation in Kombination mit der Humanevaluation gibt uns die Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung bei der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS – Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS – Q. vor KS) zusammenhängt. Mithilfe des Spearman-Tests erwies sich (Tabelle 5.56) ein signifikanter starker negativer Zusammenhang zwischen der Differenz in Fehlertyp LX.3 „Wort ausgelassen“ und der Differenz in der Stilqualität. Außerdem erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in Fehlertyp GR.10 „Falsche Wortstellung“ und der Differenz in der Stilqualität; sowie ein signifikanter schwacher

negativer Zusammenhang zwischen der Differenz in Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ und der Differenz in der Stilqualität.

Bezüglich der Inhaltsqualität erwies sich (Tabelle 5.56) ein signifikanter starker negativer Zusammenhang zwischen der Differenz in Fehlertyp LX.3 und der Differenz in der Inhaltsqualität; sowie ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.4 und GR.10 einzeln und der Differenz in der Inhaltsqualität. Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht nachgewiesen werden.

Tabelle 5.56: „Kondi. m. Wenn“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Differenz der Anzahl der **LX.3**	84	< ,001	− ,630
Differenz der Anzahl der **LX.4**	84	,010	− ,281
Differenz der Anzahl der **GR.10**	84	< ,001	− ,375
Differenz CQ (nach KS − vor KS)			
Differenz der Anzahl der **LX.3**	84	< ,001	− ,804
Differenz der Anzahl der **LX.4**	84	,046	− ,218
Differenz der Anzahl der **GR.10**	84	,016	− ,263
Differenz allg. Q (nach KS − vor KS)			
Differenz der Anzahl der **LX.3**	84	< ,001	− ,774
Differenz der Anzahl der **LX.4**	84	,014	− ,268
Differenz der Anzahl der **GR.10**	84	,004	− ,308

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Diese signifikanten negativen Korrelationen deuten darauf hin, dass sobald die Fehleranzahl der genannten Fehlertypen sank, die Qualität stieg. In Tabelle 5.55 wurde der grammatische Fehlertyp GR.10 „Falsche Wortstellung“ in ‚If‘

korrigiert, daraufhin stieg die Stil- und Inhaltsqualität jeweils um 0,75 Punkte auf der Likert-Skala.

In einem weiteren Beispiel wurde nur der lexikalische Fehler (LX.3) für die fehlende Übersetzung des Verbs ‚Ist' korrigiert, nachdem der Konditionalsatz mit ‚Wenn' formuliert wurde (Tabelle 5.57).

Tabelle 5.57: Beispiel 38

Vor-KS	**Ist** die Seriennummer des Gerätes bekannt, kann im Feld Seriennummer diese Nummer eingegeben werden.
HMÜ Bing	XXX The serial number of the device is known, this number can be entered in the "Serial Number" field.
Nach-KS	**Wenn** die Seriennummer des Gerätes bekannt **ist**, kann im Feld Seriennummer diese Nummer eingegeben werden.
HMÜ Bing	If the serial number of the device is known, this number can be entered in the "Serial Number" field.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Daraufhin stieg die Stilqualität um 0,38 Punkte und die Inhaltsqualität um 1,63 Punkte auf der Likert-Skala.

5.4.4.6.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.51 zeigt, stiegen die Stil- und Inhaltsqualität nach der Anwendung der KS-Regel nur bei dem HMÜ-System Bing (+ 15,6 % (SQ) bzw. + 33,5 % (CQ)). Zudem stieg nur die Inhaltsqualität bei dem SMÜ-System SDL (+ 16,5 %). Diese Zunahmen erwiesen sich als signifikant. Bei den anderen Systemen war die Veränderung, wenn überhaupt, minimal.

Wie die Aufteilung der Annotationsgruppen zeigte, waren 92 % der Übersetzungen vom NMÜ-System Google Translate sowohl vor als auch nach der Anwendung der KS-Regel richtig (Annotationsgruppe RR). Entsprechend ist zu erwarten, dass das Qualitätsniveau sich nicht verändert. Ebenfalls waren bei dem RBMÜ-System Lucy 71 % der MÜ in der Annotationsgruppe RR und 21 % in der Annotationsgruppe FF enthalten. Bei zwei fehlerfreien MÜ (Gruppe RR) bzw. zwei fehlerhaften MÜ (Gruppe FF) waren die Qualitätswerte vor und nach der Regelanwendung vergleichbar, daher ist die minimale Qualitätsveränderung bei

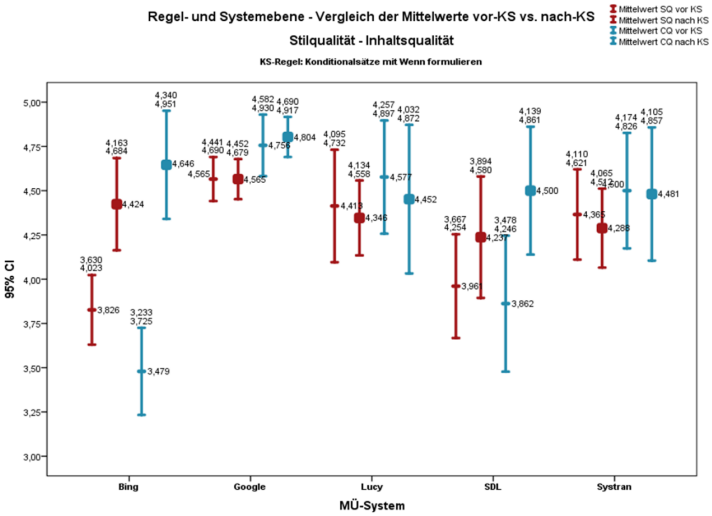


Abbildung 5.51: „Kondi. m. Wenn“ – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Lucy begründet. Bei dem HMÜ-System Systran waren die Fehleranzahl vor und nach der Anwendung der KS-Regel vergleichbar hoch und die Fehlertypen sehr gemischt, daher konnte keine signifikante Veränderung in der Qualität verzeichnet werden.

5.4.4.6.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwiesen sich bei zwei MÜ-Systemen signifikante starke negative Korrelationen (Tabelle 5.59): Bei dem HMÜ-System Bing konnte eine signifikante starke negative Korrelation zwischen der Differenz in Fehlertyp LX.3 „Wort ausgelassen“ und der Differenz der Stil- und Inhaltsqualität nachgewiesen werden, wie in Tabelle 5.57. Bei dem SMÜ-System SDL erwies sich eine signifikante starke negative Korrelation zwischen den Fehlertypen LX.3 „Wort ausgelassen“ und GR.10 „Falsche Wortstellung“ einzeln und der Stilqualität; sowie eine signifikante starke negative Korrelation zwischen LX.3 und der Inhaltsqualität.

In Tabelle 5.60 wurden die Fehlertypen LX.3 „Wort ausgelassen“ (in ‚if‘) und GR.10 „Falsche Wortstellung“ (in ‚is‘) eliminiert.

Daraufhin stieg die Stilqualität um 1,25 Punkte und die Inhaltsqualität um 0,75 Punkte auf der Likert-Skala.

Tabelle 5.58: „Kondi. m. Wenn“ – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	18	,002	− 3,108	18	< ,001	− 3,489	18	,001	− 3,422
Google	21	,982	− ,022	21	,353	− ,929	21	,516	− ,649
Lucy	13	,328	− ,978	13	,168	− 1,378	13	,027	− 2,209
SDL	19	,080	− 1,754	19	,011	− 2,528	19	,008	− 2,637
Systran	13	,348	− ,938	13	,797	− ,257	13	,366	− ,904

p: Signifikanz z: Teststatistik nicht signifikant (p ≥ 0,05)

5.4.4.7 Vergleich der MÜ-Qualität bei Konditionalsätzen mit vs. ohne ‚Wenn‘ auf Annotationsgruppenebene

Mit Ausnahme der Gruppe FR fiel die Differenz in der Stil- und Inhaltsqualität[34] bei allen anderen Annotationsgruppen sehr gering aus (Abbildung 5.52).

In der Gruppe FF (Übersetzung vor und nach KS falsch) blieb die Inhaltsqualität fast unverändert und die Stilqualität sank minimal (Abbildung 5.52). Es gab in dieser Gruppe keinen bestimmten Fehler, der vor oder nach der Anwendung der KS auftrat bzw. eliminiert wurde.

In der Gruppe FR stiegen erwartungsgemäß die Stil- und Inhaltsqualität signifikant (Abbildung 5.52): Ein Anstieg von 18,6 % bei der Stilqualität (z (N = 23) = − 4,143 / p < ,001) bzw. ein doppelt so hoher Anstieg von 36,3 % bei der Inhaltsqualität (z (N = 23) = − 4,209 / p < ,001) (Tabelle 5.61). Die Eliminierung des Fehlertyps LX.3 (das Auslassen von ‚If‘) führte eindeutig zu einer hohen Verständlichkeit und Genauigkeit der MÜ (hohe Inhaltsqualität) und steigerte die stilistische Adäquatheit[35] für eine Formulierung als Bedingung. Dieses Ergebnis wird in Tabelle 5.62 verdeutlicht, in dem die Stilqualität um 0,88 und die Inhaltsqualität um 1,63 Punkte auf der Likert-Skala stiegen.

[34] Definitionen der Qualität unter §4.5.5.1.

[35] Stilistische Adäquatheit im Sinne von Hutchins & Somers (1992: 163) ist „the extent to which the translation uses the language appropriate to its content and intention”. Mehr zu den Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.59: „Kondi. m. Wenn“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Bing			SDL		
	N	p	ρ	N	p	ρ
Differenz SQ (nach KS − vor KS)						
Diff. der Anzahl **LX.3 „W. fehlt“**	18	,029	− ,513	19	,005	− ,611
Diff. der Anzahl **GR.10 „Wortst.“**				19	,008	− ,588
Differenz CQ (nach KS − vor KS)						
Diff. der Anzahl **LX.3 „W. fehlt“**	18	,019	− ,548	19	< ,001	− ,757
Diff. der Anzahl **GR.10 „Wortst.“**				19	,306	− ,248
Differenz Q (nach KS − vor KS)						
Diff. der Anzahl **LX.3 „W. fehlt“**	18	,019	− ,547	19	< ,001	− ,766
Diff. der Anzahl **GR.10 „Wortst.“**				19	,058	− ,443

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (ρ ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Die Gruppe RF war, wie die Aufteilung der Annotationsgruppen (§5.4.4.3) zeigte, sehr selten vertreten (nur 5 % der Fälle) (Abbildung 5.44). Übersetzungen dieser Gruppe kamen in grenzwertigen Fällen vor, in denen z. B. die Konditionalkonjunktion ‚Wenn‘ mangels Kontextinformationen als ‚When‘ anstatt ‚If‘ übersetzt wurde (siehe Tabelle 5.54). In der Gruppe RR (Übersetzung vor und nach KS richtig) waren die Übersetzungen vor und nach der Anwendung der KS-Regel identisch, so dass es keinen Raum für eine Veränderung im Qualitätsniveau gibt.

Tabelle 5.60: Beispiel 39

Vor-KS	**Steht** ein normierter Faktor zur Verfügung, kann dieser Faktor direkt in der Eingabemaske eingegeben werden.
SMÜ SDL	XXX Is a standardized factor available, this factor can be entered directly in the input mask.
Nach-KS	**Wenn** ein normierter Faktor zur Verfügung **steht**, kann dieser Faktor direkt in der Eingabemaske eingegeben werden.
SMÜ SDL	If a standardized factor is available, this factor can be entered directly in the input mask.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

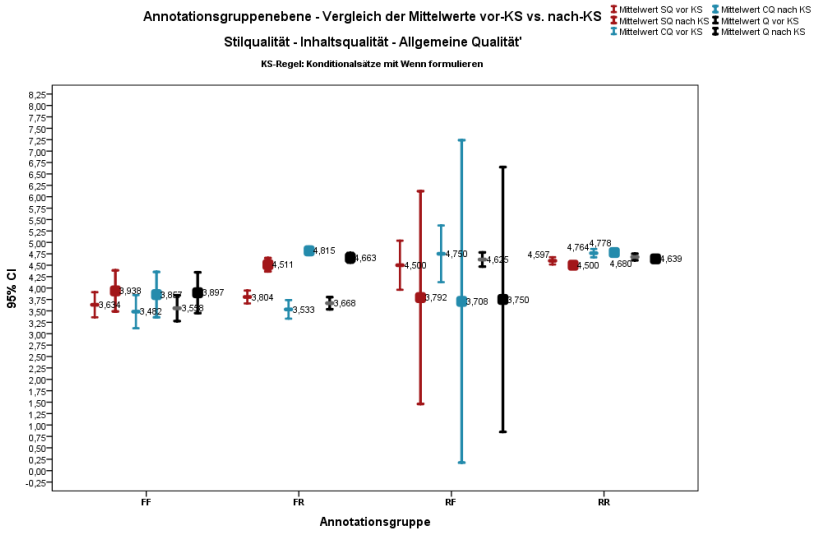


Abbildung 5.52: „Kondi. m. Wenn“ – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

Tabelle 5.61: „Kondi. m. Wenn“ – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	14	,090	− 1,694
Differenz CQ (nach KS − vor KS)	14	,278	− 1,084
Differenz allg. Q (nach KS − vor KS)	14	,157	− 1,414
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	23	< ,001	− 4,143
Differenz CQ (nach KS − vor KS)	23	< ,001	− 4,209
Differenz allg. Q (nach KS − vor KS)	23	< ,001	− 4,201
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	3	,285	− 1,069
Differenz CQ (nach KS − vor KS)	3	,285	− 1,069
Differenz allg. Q (nach KS − vor KS)	3	,285	− 1,069
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	44	,620	− ,222
Differenz CQ (nach KS − vor KS)	44	,795	− ,260
Differenz allg. Q (nach KS − vor KS)	44	,229	− 1,204

Tabelle 5.62: Beispiel 40

Vor-KS	**Ist** ein mehrstufiges Modul **parametriert**, so sind die externen Kontakte zu verriegeln.
HMÜ Bing	XXX A multi-level module is programmed, the external contacts must be locked.
Nach-KS	**Wenn** ein mehrstufiges Modul **parametriert ist**, sind die externen Kontakte zu verriegeln.
HMÜ Bing	If a multi-level module is programmed, the external contacts must be locked.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

5.4.4.8 Vergleich der AEM-Scores bei Konditionalsätzen mit vs. ohne ‚Wenn' sowie die Korrelation zwischen den AEM-Scores und der Qualität

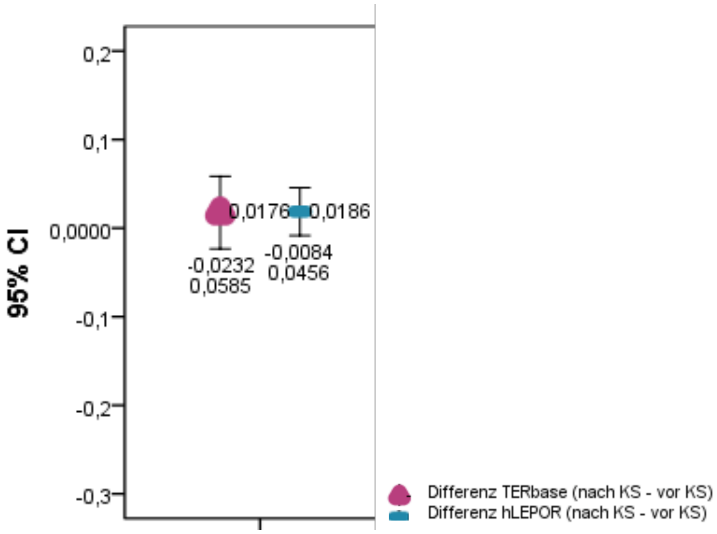
Der Vergleich der AEM-Scores vor und nach der Formulierung der Konditionalsätze mit ‚Wenn' zeigte sowohl mit TERbase als auch mit hLEPOR nur eine geringe nicht signifikante Verbesserung der AEM-Scores (Abbildung 5.53).

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei ,018 (SD = ,188) und für die hLEPOR bei ,019 (SD = ,124) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben) (Abbildung 5.53). Durch diese minimalen Unterschiede waren die Differenzen (nach KS − vor KS) in TERbase und hLEPOR nicht signifikant (z (N = 84) = − 1,223 / p = ,221) bzw. (z (N = 84) = − 1,774 / p = ,076). Wie der Vergleich der Fehlertypen (siehe §5.4.4.4) zeigt, war der dominante Fehlertyp (LX.3) das Fehlen der Konjunktion ‚If' vor der Anwendung der Regel. In Tabelle 5.63 wird der Fehler LX.3 (vor KS) durch das Einfügen von ‚If' (nach KS) korrigiert.

Außer diesem Fehler (Fehlen der Konjunktion ‚If') waren viele MÜ-Sätze – vergleichbar mit dem Tabelle 5.34 – in beiden Szenarien identisch. Das kann der Grund für die kleine (insignifikante) Differenz in den AEM-Scores (nach KS − vor KS) gewesen sein.

5.4.4.8.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter starker positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.53: „Kondi. m. Wenn“ – Mittelwert der Differenz der AEM-Scores

Tabelle 5.63: Beispiel 41

Vor-KS	**Liegt** die Regelabweichung innerhalb der x-Zone, so bleibt das erste Modul stehen.
HMÜ Bing	XXX The control deviation is within the x-zone, the first module remains stationary.
Nach-KS	**Wenn** die Regelabweichung innerhalb der x-Zone **liegt**, bleibt das erste Modul stehen.
HMÜ Bing	If the control deviation is within the x-zone, the first module remains stationary.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

und hLEPOR und der Differenz in der allgemeinen Qualität. Tabelle 5.64 demonstriert die Korrelationswerte.

Tabelle 5.64: „Kondi. m. Wenn“ – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS – vor KS)	84	< ,001	,551	starker Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS – vor KS)	84	< ,001	,582	starker Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Dieses Ergebnis weist darauf hin, dass – nach der Formulierung der Konditionalsätze mit ‚Wenn‘ – der Anstieg der Qualität mit einer Verbesserung der Scores der beiden AEMs einherging.

5.4.4.9 Analyse der dritten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der dritten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (–) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[36]

[36] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Regel 3: Konditionalsätze mit ‚Wenn' einleiten

Erster Analysefaktor: Vergleich der Fehleranzahl bei Konditionalsätzen mit vs. ohne ‚Wenn'

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Die Fehleranzahl sank signifikant, nachdem die Konditionalsätze mit ‚Wenn' formuliert wurden. **Anz.F. (–)**

Auf Regel- und MÜ-Systemebene:
Bei Bing und SDL sank die Fehleranzahl signifikant, nachdem die Konditionalsätze mit ‚Wenn' formuliert wurden. **Bi (–)** **SD (–)**

Bei Lucy und Systran sank ebenfalls die Fehleranzahl, jedoch war der Rückgang nicht signifikant. **Lu (–)** **Sy (–)**

In Google war die Fehleranzahl sehr gering: 1 Fehler vor KS und 2 Fehler nach KS. **Go (+)**

Zweiter Analysefaktor

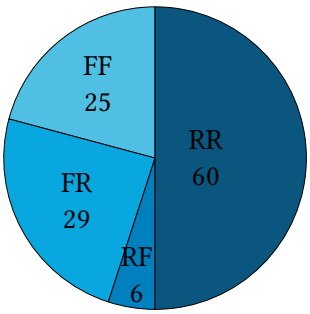


Abbildung 5.54: Aufteilung der Annotationsgruppen auf Regelebene

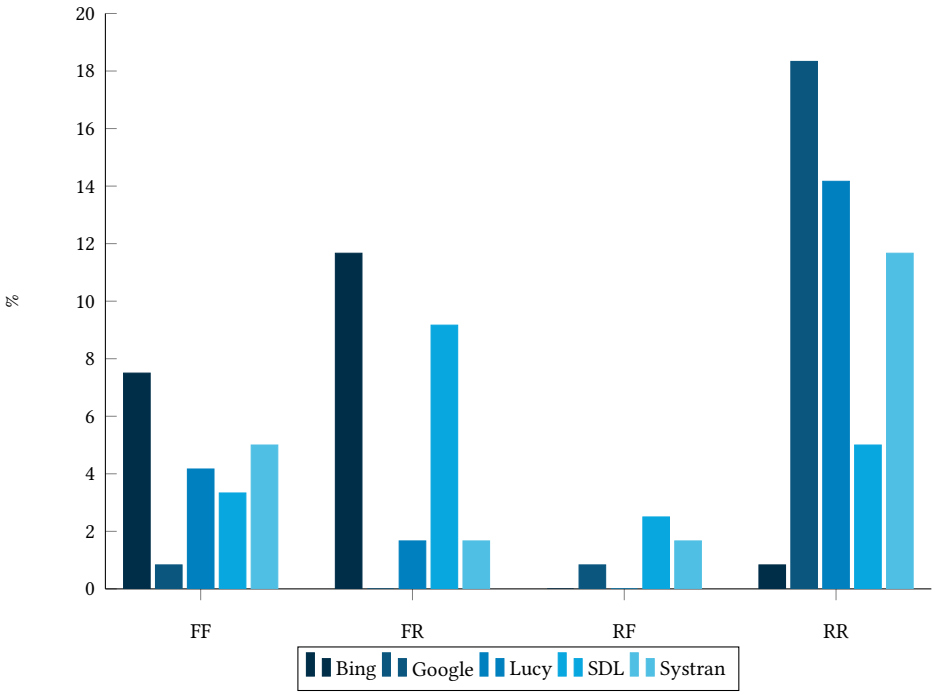


Abbildung 5.55: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

---

Dritter Analysefaktor: Vergleich der Fehlertypen bei Konditionalsätzen mit vs. ohne ‚Wenn'

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H1 wurde nur für zwei Fehlertypen bestätigt.
Die Fehleranzahl von LX.3 „Wort ausgelassen" und LX.4 „Zusätzliches Wort eingefügt" sanken signifikant, nachdem die Konditionalsätze mit ‚Wenn' formuliert wurden.

LX.3 (–)
LX.4 (–)

Auf Regel- und MÜ-Systemebene:
Nachdem die Konditionalsätze mit ‚Wenn' formuliert wurden, sank **die** Fehleranzahl von LX.3 „Wort ausgelassen" bei Bing und SDL und von GR.10 „Falsche Wortstellung" bei Bing signifikant.

LX.3 (−): **Bi SD**
GR.10 (−): **Bi**

Alle weiteren Veränderungen waren nicht signifikant.

Vierter Analysefaktor: Vergleich der MÜ-Qualität bei Konditionalsätzen mit vs. ohne ‚Wenn'

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Sowohl die Stil- als auch die Inhaltsqualität stiegen signifikant.

SQ (+)
CQ (+)

Auf Regel- und MÜ-Systemebene:
Die Stilqualität stieg nur bei Bing signifikant, während die Inhaltsqualität bei Bing und SDL signifikant stieg.

SQ (+): **Bi**
CQ (+): **Bi SD**

Die Qualitätsdifferenzen bei allen anderen Systemen waren (sehr) klein und entsprechend nicht signifikant.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde nur für zwei Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“ und der Differenz der Stilqualität und Inhaltsqualität.
Zudem bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Falsche Wortstellung“ und der Differenz der Stilqualität.

***neg* LX.3 <<>> SQ**
***neg* GR.10 <> SQ**
***neg* LX.3 <<>> CQ**

Auf Regel- und MÜ-Systemebene:
Bei Bing bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“ und der Differenz der Stil- und Inhaltsqualität.

Bi
***neg* LX.3 <<>> SQ**
***neg* LX.3 <<>> CQ**

Bei SDL besteht ebenfalls ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“ und der Differenz der Stil- und Inhaltsqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Falsche Wortstellung“ und der Differenz der Stilqualität.

SD
***neg* LX.3 <<>> SQ**
***neg* LX.3 <<>> CQ**
***neg* GR.10 <<>> SQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität bei Konditionalsätzen mit vs. ohne ‚Wenn‘ auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde nur für die Gruppe FR bestätigt: Bei der Annotationsgruppe FR stiegen die Stil- und Inhaltsqualität signifikant, nachdem die Konditionalsätze mit ‚Wenn' formuliert wurden.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe FF stiegen die Stil- und Inhaltsqualität leicht.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe RR sank die Stilqualität leicht und die Inhaltsqualität stieg minimal.	**SQ (–)** **CQ (+)**
Bei der Annotationsgruppe RF sanken die Stil- und Inhaltsqualität insignifikant.	**SQ (–)** **CQ (–)**

Siebter Analysefaktor: Vergleich der AEM-Scores bei Konditionalsätzen mit vs. ohne ‚Wenn'

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. Die AEM-Scores von TERbase und hLEPOR stiegen nur leicht nachdem die Konditionalsätze mit ‚Wenn' formuliert wurden.	**TERbase (+)** **hLEPOR (+)**

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter starker positiver Zusammenhang zwischen den Differenzen der Scores der beiden AEMs (TERbase und hLEPOR) und der Differenz der allgemeinen Qualität.

***pos* TERbase <<>> Q**
***pos* hLEPOR <<>> Q**

5.4.5 VIERTE REGEL: Eindeutige pronominale Bezüge verwenden

5.4.5.1 Überblick

Im Folgenden wird die KS-Regel „Eindeutige pronominale Bezüge verwenden" kurz beschrieben.[37] Zudem wird zusammenfassend und anhand von Beispielen demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Eindeutige pronominale Bezüge verwenden (tekom-Regel-Nr. S 102)

Nach dieser Regel (tekom 2013: 60) sollen die Pronomen vermieden werden, wenn sie mehrdeutig sein könnten. Anstelle des Pronomens soll das Bezugswort wiederholt werden, damit der Bezug eindeutig erkennbar wird.

Begründung: Der Leser eines technischen Dokuments ist in der Regel weniger mit der Thematik vertraut als der Autor (ebd.: 61).

[37] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

Umsetzungsmuster:

Vor KS: Der Satz beinhaltet ein Pronomen.
Nach KS: Das Pronomen wird durch das Nomen ersetzt.

KS-Stelle

Vor KS: das Pronomen (Personalpronomen und Demonstrativpronomen)
Nach KS: das Nomen bzw. das Demonstrativpronomen und das Nomen (inkl. damit verbundener Fehler in der Wortstellung)

Beispiele

Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, ihn rückstandslos zu entfernen.

Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, den Fleck rückstandslos zu entfernen.

Sofern auf der Oberfläche alte Kleberreste anhaften, sind diese vollständig zu entfernen.

Sofern auf der Oberfläche alte Kleberreste anhaften, sind diese Kleberreste vollständig zu entfernen.

Aufteilung der Testsätze: Die im Datensatz abgedeckten Pronomen sind 9 Personalpronomen in verschiedenen Kasus und 15 Demonstrativpronomen. Diese Verteilung der beiden Pronomenarten spiegelt deren Verteilung im Korpus wider, denn die Demonstrativpronomen kommen im Vergleich zu den Personalpronomen häufiger vor.

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.5.2 Vergleich der Fehleranzahl vor vs. nach der Verwendung von pronominalen Bezügen

Die Fehleranzahl sank minimal um 9,3 % von 43 Fehlern im Falle der Verwendung von Pronomen (M = ,36 / SD =,577 / N = 120) auf 39 Fehler bei der Verwendung von pronominalen Bezügen (M = ,33 / SD = ,610 / N = 120) (Abbildung 5.56 und Abbildung 5.57). Der Mittelwert der Differenz (nach KS − vor KS) der Fehleranzahl pro Satz lag somit bei − ,03 (SD = ,697) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,16 (SD = ,577) und einem Maximum von ,09

(SD = ,803) (Bootstrapping mit 1000 Stichproben). Entsprechend war die Differenz (nach KS – vor KS) der Fehleranzahl nicht signifikant (z (N = 120) = – ,503 / p = ,615).

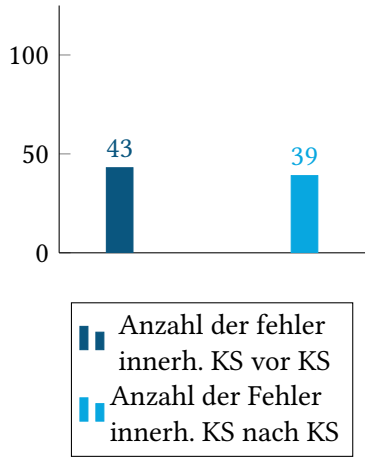


Abbildung 5.56: „Pronom. Bezüge verw." – Fehlersumme vor vs. nach KS

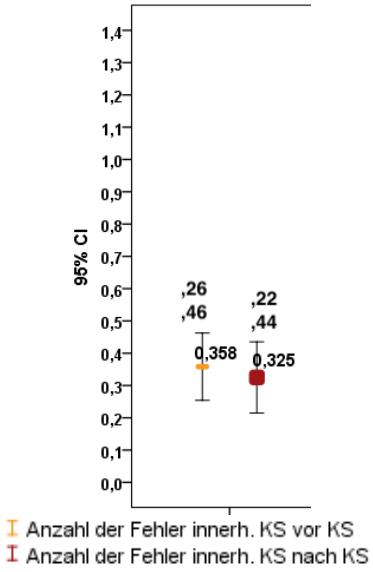


Abbildung 5.57: „Pronom. Bezüge verw." – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Ein Vergleich der Annotationsgruppen RF und FR zeigt, dass relativ viele Fehler bei der Verwendung der Demonstrativpronomen (24 %) im Vergleich zu den Personalpronomen (7 %) auftraten und mit der Umsetzung der KS-Regel eliminiert wurden (Tabelle 5.65).

Tabelle 5.65: Daten der untersuchten Pronomen

	RF	FR
Demonstrativpronomen Anz. d. Ausgangssätze 15; bei 5 MÜ-Systemen = 15 * 5 = 75 MÜ	8 MÜ (11 %)	18 MÜ (24 %)
Personalpronomen Anz. d. Ausgangssätze 9; bei 5 MÜ-Systemen = 9 * 5 = 45 MÜ	6 MÜ (13 %)	3 MÜ (7 %)

Wie Tabelle 5.65 zeigt, war die Annotationsgruppe FR größer als RF. Dies führte zu der allgemeinen Abnahme der Fehleranzahl nach der Regelanwendung. Dennoch war die Annotationsgruppe RF relativ breit vertreten, daher ist die Differenz in der Fehleranzahl nicht hoch. Unter §5.4.5.3 werden die Annotationsgruppen unter die Lupe genommen.

5.4.5.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Eine genauere Betrachtung der einzelnen MÜ-Systeme zeigt, dass die Systeme sehr unterschiedlich auf die KS-Regel reagierten:

Die Fehleranzahl bei dem NMÜ-System Google Translate stieg signifikant Mdiff = 0,167 (z (N = 24) = − 2,000 / p = ,046), allerdings wäre eine bloße Betrachtung der quantitativen Daten in diesem Fall irreführend, denn Google Translate konnte in 83 % der Fälle (20 von 24 Sätzen) eine korrekte MÜ sowohl für das Pronomen (vor KS) als auch für den pronominalen Bezug (nach KS) liefern (siehe §5.4.5.3). Nur in 17 % der Sätze (4 von 24) stieg die Fehleranzahl (insgesamt +4 Fehler).

Auf der anderen Seite blieb die Fehleranzahl bei dem HMÜ-System Bing unverändert und sank bei den weiteren drei Systemen: RBMÜ-System Lucy (Mdiff = − ,125); HMÜ-System Systran (Mdiff = − ,125); SMÜ-System SDL (Mdiff = − ,083). Dieser Rückgang erwies sich als insignifikant.

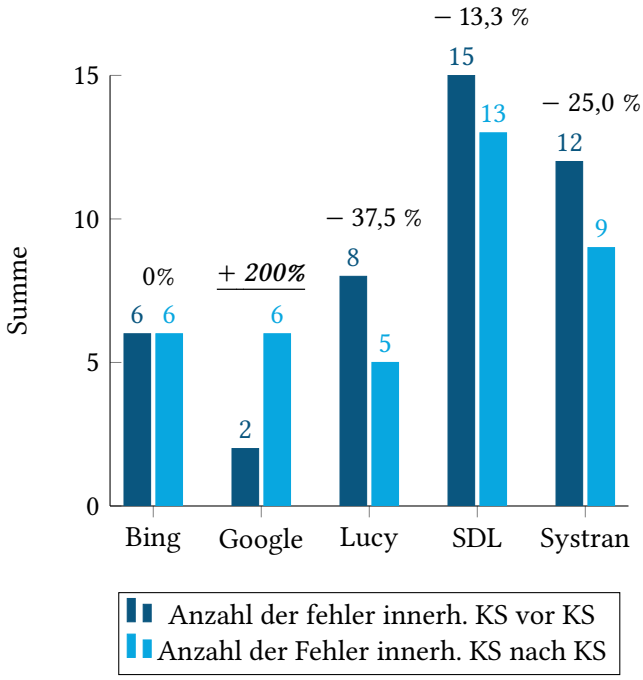
Tabelle 5.66 zeigt einen der wenigen Fälle, in denen Google Translate nur bei der Verwendung eines pronominalen Bezugs (nach KS) eine falsche Übersetzung lieferte, wobei Lucy (siehe Tabelle 5.68) und Bing (siehe Tabelle 5.69) im selben Beispielsatz das Pronomen (vor KS) falsch und den pronominalen Bezug (nach KS) richtig übersetzten.

In Tabelle 5.66 handelt es sich um einen Wortstellungsfehler. Mehr zu den Fehlertypen und potenziellen Gründen dahinter wird unter §5.4.5.4 diskutiert.

5.4.5.3 Aufteilung der Annotationsgruppen

Wie Abbildung 5.59 veranschaulicht, war der Anteil der Annotationsgruppe RR besonders hoch (57,5 %). Bei 17,5 % der Sätze war die MÜ bei der Verwendung des Pronomens falsch und wurde nach der Verwendung des pronominalen Bezugs korrigiert (Gruppe FR).

Die beiden weiteren Annotationsgruppen (FF und RF) waren ähnlich klein. Ein Anteil von nur 11,7 % bei der Gruppe RF besagt, dass eine Falschübersetzung der Sätze erst nach der Anwendung der KS-Regel eine relative Seltenheit war.



Signifikante Differenz vor vs. nach KS

Abbildung 5.58: „Pronom. Bezüge verw." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.66: Beispiel 42

Vor-KS	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **diese** warten.
GNMÜ	Only qualified electricians are allowed to access to the machine's electrical system and maintain it.
Nach-KS	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **die Maschine** warten.
GNMÜ	Only qualified electricians are allowed to access to the machine's electrical system and the machine maintain.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

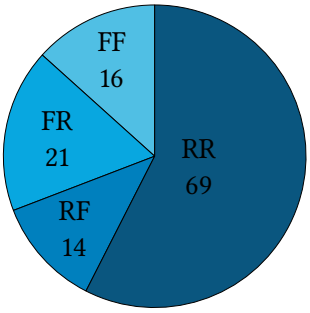


Abbildung 5.59: „Pronom. Bezüge verw." – Aufteilung der Annotationsgruppen

5.4.5.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

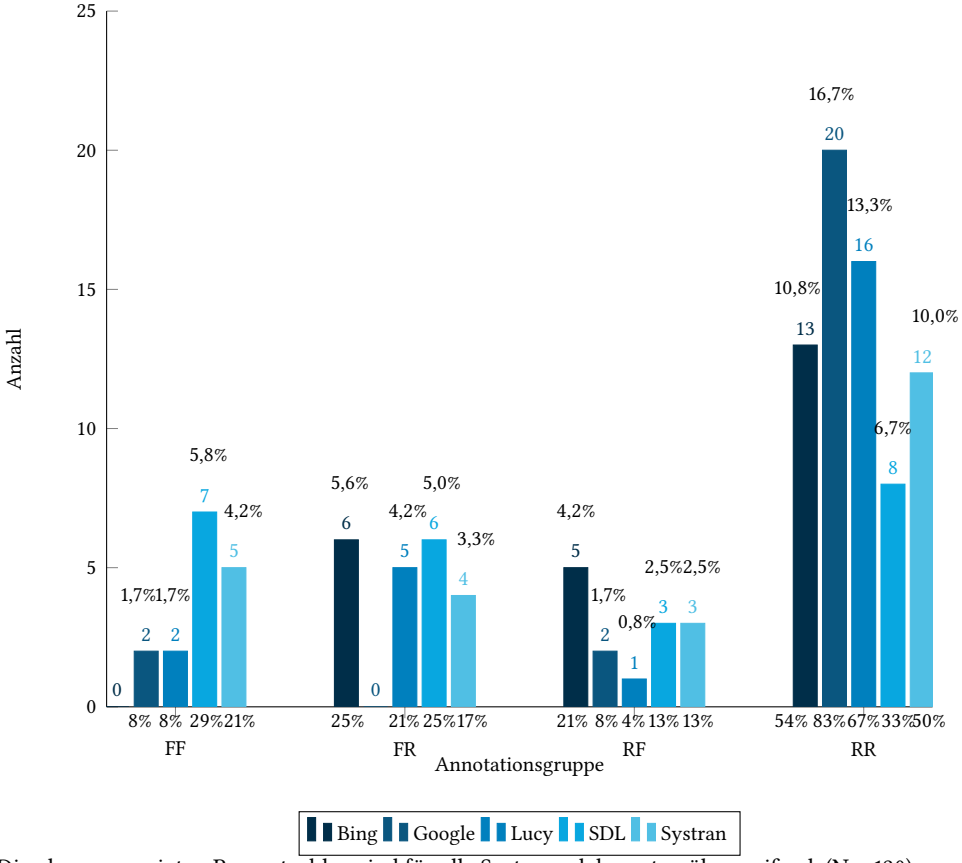
Der dominante Anteil der Annotationsgruppe RR (richtig sowohl vor als auch nach der Anwendung der KS-Regel) ist bei allen MÜ-Systemen ohne Ausnahme zu sehen. Eine genaue Betrachtung der Gruppe RR (Abbildung 5.60) auf MÜ-Systemebene zeigt, dass das NMÜ-System Google Translate mit 83 % seiner Übersetzungen auf den ersten Platz kommt; gefolgt vom RBMÜ-System Lucy mit 67 %; dem HMÜ-System Bing mit 54 %; dem HMÜ-System Systran mit 50 % und zum Schluss mit dem SMÜ-System SDL mit 33 %.

SDL verzeichnete den niedrigsten Anteil bei der Gruppe RR und gleichzeitig den höchsten Anteil bei der Gruppe FF (Abbildung 5.60). Da dieses System rein statistisch arbeitet, zeigt dieses Ergebnis, dass ein Großteil der Testsätze selten bzw. nicht in den Trainingsdaten vorkommt.

5.4.5.4 Vergleich der Fehlertypen vor vs. nach der Verwendung von pronominalen Bezügen

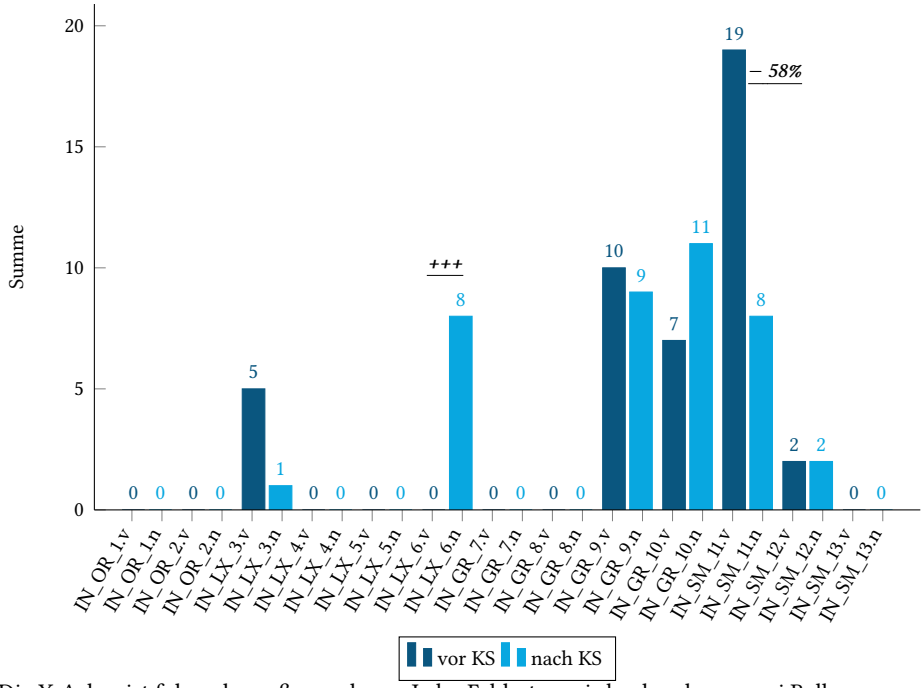
Nach der Verwendung von pronominalen Bezügen verändert sich die Anzahl zweier Fehlertypen deutlich, und zwar stieg die Fehleranzahl bei Fehlertyp LX.6 „Lexik – Konsistenzfehler" und sank bei Fehlertyp SM.11 „Semantik – Verwechslung des Sinns". Beide Veränderungen erwiesen sich als signifikant p = ,013 beim LX.6 bzw. p = ,027 beim SM.11 (N = 120).

Bei der Umsetzung dieser KS-Regel wird anstatt des Pronomens, der Pronominalbezug verwendet, entsprechend wird oft ein Nomen, das z. B. im Hauptsatz vorkommt, im Nebensatz wiederholt. Die Systeme übersetzten jedoch die zweite Instanz des Nomens in manchen Fällen anders, wodurch der Fehlertyp LX.6



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.60: „Pronom. Bezüge verw." – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler *vor KS* und der zweite die Summe der Fehler *nach KS*, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
__*Signifikante Differenz vor vs. nach KS*__
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehle

Abbildung 5.61: „Pronom. Bezüge verw.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

„Lexik – Konsistenzfehler“ nach der Anwendung der KS-Regel (Mv = – / SDv = – / Mn = ,07 / SDn = ,250 / N = 120) auftrat. Tabelle 5.67 veranschaulicht diese Problematik.

Tabelle 5.67: Beispiel 43

Vor-KS	Fettreste müssen vollständig abgewaschen werden, da sich **diese** ansonsten in der Pfanne einbrennen können.
SMÜ SDL	Grease residue must be completely washed off as it can otherwise burn in the pan.
Nach-KS	Fettreste müssen vollständig abgewaschen werden, da sich **diese Reste** ansonsten in der Pfanne einbrennen können.
SMÜ SDL	Grease residue must be completely washed off as this remains can otherwise burn in the pan.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

In Tabelle 5.67 konnte das System das Pronomen (vor KS) problemlos übersetzen, während der Pronominalbezug (nach KS) abweichend übersetzt wurde. In der Regel führen die Unternehmen Terminologiedatenbanken und integrieren sie in ihre implementierten MÜ-Systeme, so dass ein Konsistenzfehler dieser Art (vgl. Mertin 2006: 249) nur bei neuen bzw. nicht eingepflegten Termini vorkommen würde.

Auf der anderen Seite entstand bei der Verwendung von Pronomen (vor KS) ein semantisches Problem, nämlich die Verwechslung ihres Sinns; ein Fehler, der – trotz der Forschungsfortschritte – auf der weiterhin bestehenden Schwierigkeit einer Koreferenzauflösung (d. h. Identifizierung der Entität, auf die sich die Koreferenz bzw. das Pronomen bezieht) beruht (Ng 2017). Dieser Fehler kam insbesondere bei der Übersetzung von Demonstrativpronomen (‚dies‘ und ‚diese‘) vor. Nach der Verwendung des Pronominalbezugs sank entsprechend die Fehleranzahl des Fehlertyps SM.11 „Semantik – Verwechslung des Sinns“ von 19 auf 8 Fehler (- 57,9 % / Mv = ,16 / SDv = ,367 / Mn = ,07 / SDn = ,250 / N = 120).

In Tabelle 5.68 wurde der Sinn des Demonstrativpronomens ‚diese‘ vom System nicht richtig erkannt. Auf der anderen Seite wurde der Pronominalbezug korrekt übersetzt.

Tabelle 5.68: Beispiel 44

Vor-KS	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **diese** warten.
RBMÜ Lucy	Only qualified electricians are allowed to access to the machine's electrical system and maintain these.
Nach-KS	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **die Maschine** warten.
RBMÜ Lucy	Only qualified electricians are allowed to access to the machine's electrical system and maintain the machine.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.5.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

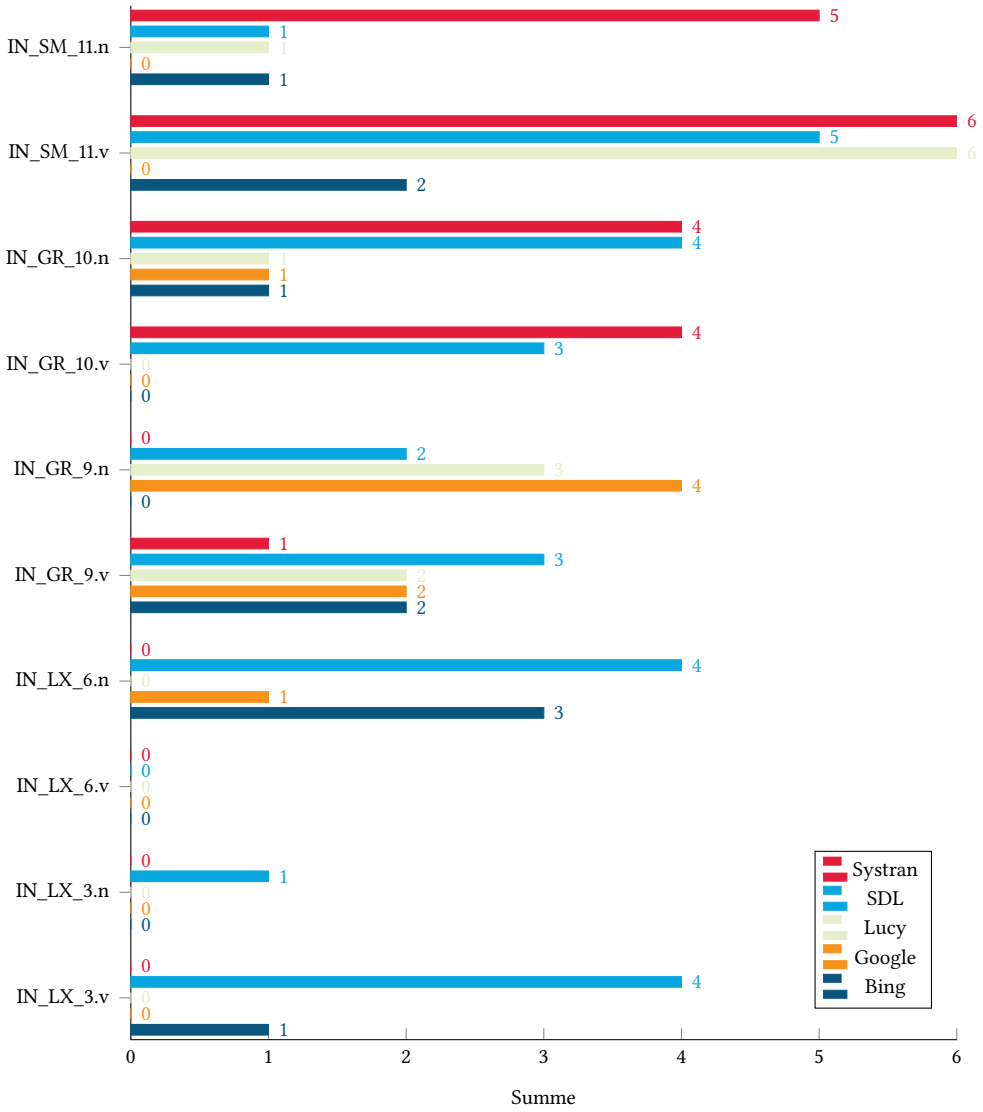
Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt, dass gar kein Fehlertyp sich bei einem bestimmten System signifikant veränderte. Auch die Differenzen des Fehlertyps LX.6 „Lexik – Konsistenzfehler" und Fehlertyp SM.11 „Semantik – Verwechslung des Sinns" waren bei keinem der analysierten Systeme signifikant.

Wesentliche (nicht signifikante) Veränderungen gab es jedoch bei zwei Systemen: Bei SDL stieg die Fehleranzahl vom Fehlertyp LX.6 „Lexik – Konsistenzfehler" von 0 auf 4 Fehler. Dies kann bei einem rein SMÜ-System auf das Nichtvorhandensein der Segmente in den Trainingsdaten zurückgeführt werden. Bei Lucy nahm die Fehleranzahl des Fehlertyps SM.11 „Semantik – Verwechslung des Sinns" von 6 auf 1 Fehler ab. Hierbei zeigte das RBMÜ-System eine Schwäche beim Übersetzen von Demonstrativpronomen, denn alle fünf korrigierten Fehler der Demonstrativpronomen waren für das System ambig.

5.4.5.5 Vergleich der MÜ-Qualität vor vs. nach der Verwendung von pronominalen Bezügen sowie die Korrelation zwischen den Fehlertypen und der Qualität

Nach der Verwendung der pronominalen Bezüge gab es einen kleinen insignifikanten Rückgang bei der Stilqualität gekoppelt mit einem kleinen insignifikanten Anstieg bei der Inhaltsqualität (Abbildung 5.63):[38]

[38] Definitionen der Qualität unter §4.5.5.1.



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 1, 2, 4, 5, 7, 8 und 13 bei gar keinem MÜ-System vor. Zudem kam der Fehlertyp 12 nur einmal jeweils bei 3 MÜ-Systemen vor.

Abbildung 5.62: „Pronom. Bezüge verw.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

Die Stilqualität sank um 1,5 % (Mv = 4,04 / SDv = ,495 / Mn = 3,98 / SDn = ,401 / N = 77). Die Inhaltsqualität stieg um 2,3 % (Mv = 4,43 / SDv = ,554 / Mn = 4,53 / SDn = ,494 / N = 77). Der Mittelwert der Differenz (nach KS – vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei – ,063 (SD = ,460) mit einem 95%-Konfidenzintervall zwischen einem Minimum von – ,168 und einem Maximum von ,041 und für die Inhaltsqualität bei ,101 (SD = ,584) mit einem 95%-Konfidenzintervall zwischen einem Minimum von – ,032 und einem Maximum von ,233 (Bootstrapping mit 1000 Stichproben) (Abbildung 5.64). Die Differenzen (nach KS – vor KS) in der Stil- und Inhaltsqualität waren nicht signifikant (z (N = 77) = – 1,333 / p = ,183) bzw. (z (N = 77) = – 1,719 / p = ,086).

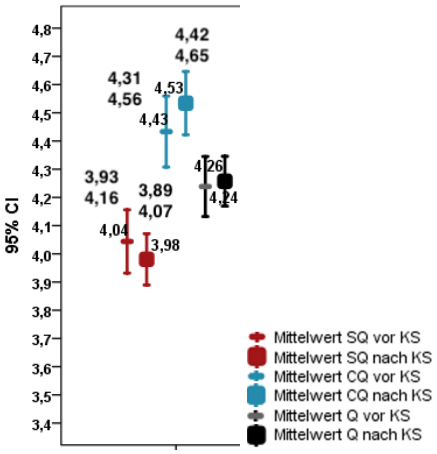


Abbildung 5.63: „Pronom. Bezüge verw." – Mittelwerte der Qualität vor und nach KS

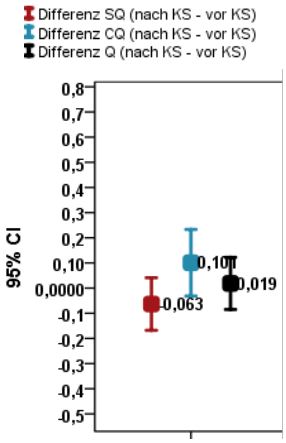


Abbildung 5.64: „Pronom. Bezüge verw." – Mittelwert der Qualitätsdifferenzen

Grundsätzlich ist ein solches Ergebnis vorstellbar, denn die Wiederholung des Bezugs unterstützt durch eine höhere Eindeutigkeit die MÜ. Dies wiederum sorgt für eine bessere Verständlichkeit des MÜ-Outputs. Gleichzeitig wird auf stilistischer Ebene der MÜ-Output durch die Wiederholung des Bezugs beeinträchtigt.

In Tabelle 5.69 wurde der Kongruenzfehler im Pronomen ‚them' nach der Anwendung der KS-Regel behoben. Daraufhin stieg die Inhaltsqualität der MÜ (+0,50 Punkte auf der Likert-Skala), während die Stilqualität sank (– ,13 Punkte auf der Likert-Skala).

Die Humanevaluation lieferte genauere Inputs zu den beeinflussten Qualitätskriterien. Wie die Abbildung 5.65 zeigt, sanken alle Qualitätskriterien mit Ausnahme des Stilqualitätskriteriums SQ3 „natürliche bzw. idiomatische MÜ" nach

Tabelle 5.69: Beispiel 45

Vor-KS	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **diese** warten.
HMÜ Bing	Only qualified electricians are allowed to access to the machine's electrical system and maintain them.
Nach-KS	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **die Maschine** warten.
HMÜ Bing	Only qualified electricians are allowed to access to the machine's electrical system and maintain the machine.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

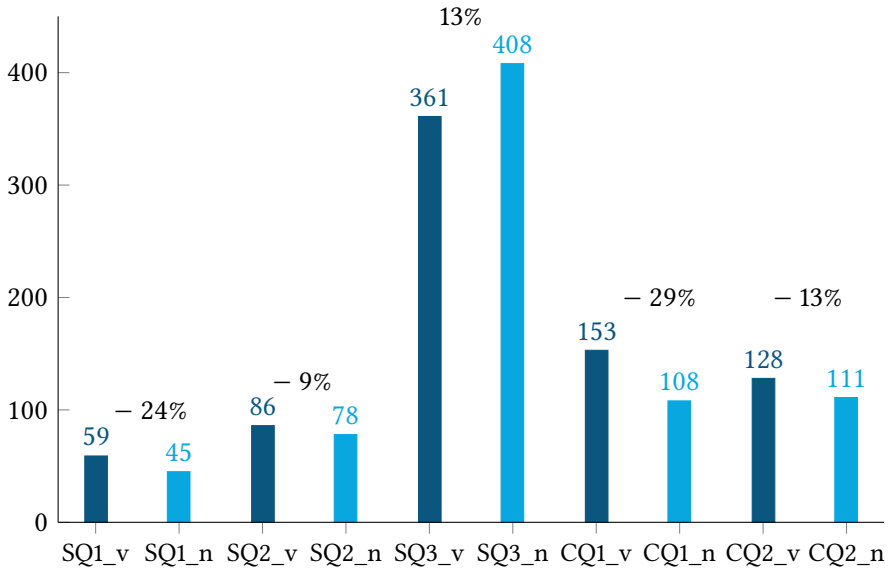
der Regelanwendung leicht. Da die Qualitätskriterien negativ formuliert sind, deutet ein Rückgang auf eine Verbesserung der Qualität hin (d. h. alle Qualitätskriterien mit Ausnahme der Idiomatik (SQ3) wurden nach der Regelanwendung besser bewertet).

An dieser Stelle folgt Tabelle 5.70, in dem die MÜ vor und nach der Anwendung der Regel fehlerfrei war, jedoch ging die Verwendung des pronominalen Bezugs mit einem Anstieg der Inhaltsqualität (+ 0,50 Punkte auf der Likert-Skala) und einem Rückgang der Stilqualität (− ,88 Punkte auf der Likert-Skala) einher.

Tabelle 5.70: Beispiel 46

Vor-KS	Um die Startlinie festzustellen, kann mit Kreide **diese** markiert werden.
HMÜ Bing	In order to determine the starting line, it can be marked with chalk.
Nach-KS	Um die Startlinie festzustellen, kann mit Kreide **diese Linie** markiert werden.
HMÜ Bing	In order to determine the starting line, this line can be marked with chalk.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder.
CQ2: Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.65: „Pronom. Bezüge verw." – Vergleich der Qualitätskriterien

5.4.5.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung bei der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS – Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS – Q. vor KS) zusammenhängt. Mithilfe des Spearman-Tests erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in den Fehlertypen GR.10 „Falsche Wortstellung" und SM.11 „Verwechslung des Sinns" einzeln und der Differenz in der Stilqualität; sowie ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz im Fehlertyp LX.6 „Konsistenzfehler" und der Differenz in der Stilqualität. (Tabelle 5.71)

Bezüglich der Inhaltsqualität erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.6, GR.10 und SM.11 einzeln und der Differenz in der Inhaltsqualität; sowie ein signifikanter

schwacher negativer Zusammenhang zwischen der Differenz in Fehlertyp LX.3 „Wort ausgelassen“ und der Differenz in der Inhaltsqualität. (Tabelle 5.71)

Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht nachgewiesen werden.

Tabelle 5.71: „Pronom. Bezüge verw.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Differenz der Anzahl der **LX.3 „W. ausgelassen“**	77	,084	− ,198
Differenz der Anzahl der **LX.6 „Konsistenzfehler“**	77	,042	− ,232
Differenz der Anzahl der **GR.10 „f. Wortstellung“**	77	,003	− ,330
Differenz der Anzahl der **SM.11 „Verwechs. des Sinns“**	77	< ,001	− ,394
Differenz CQ (nach KS − vor KS)			
Differenz der Anzahl der **LX.3 „W. ausgelassen“**	77	,025	− ,256
Differenz der Anzahl der **LX.6 „Konsistenzfehler“**	77	,006	− ,309
Differenz der Anzahl der **GR.10 „f. Wortstellung“**	77	,004	− ,327
Differenz der Anzahl der **SM.11 „Verwechs. des Sinns“**	77	,004	− ,323
Differenz allg. Q (nach KS − vor KS)			
Differenz der Anzahl der **LX.3 „W. ausgelassen“**	77	,021	− ,263
Differenz der Anzahl der **LX.6 „Konsistenzfehler“**	77	,009	− ,294
Differenz der Anzahl der **GR.10 „f. Wortstellung“**	77	,001	− ,371
Differenz der Anzahl der **SM.11 „Verwechs. des Sinns“**	77	< ,001	− ,388

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz	nicht signifikant (p ≥ 0,05)	ρ: Korrelationskoeffizient
schwache Korrelation (ρ >=0,1)	mittlere Korrelation (ρ >= 0,3)	starke Korrelation (ρ >= 0,5)

Diese signifikanten negativen Korrelationen deuten darauf hin, dass sobald die Fehleranzahl eines der genannten Fehler nach der Regelanwendung sank,

die Qualität stieg und umgekehrt, bei einem Anstieg der Fehleranzahl, die Qualität sank. Tabelle 5.67 demonstriert den zweiten Fall: In Tabelle 5.67 trat bei der Wiederholung des Bezugs (nach KS) der lexikalische Fehlertyp „Konsistenzfehler“ (LX.6) in ‚remains‘ auf. Bei einer Formulierung des Satzes mit dem Pronomen (vor KS), konnte das MÜ-System (hier SDL) das Pronomen problemlos übersetzen. Daraufhin sanken die Stilqualität um 0,75 Punkte und Inhaltsqualität um 1,00 Punkte auf der Likert-Skala.

Umgekehrt wurde Tabelle 5.72 der Wortstellungsfehler (GR.10) und der semantische Fehler SM.11 in ‚this‘ nach der Verwendung des Bezugs (nach KS) korrigiert:

Tabelle 5.72: Beispiel 47

Vor-KS	Um die Startlinie festzustellen, kann mit Kreide **diese** markiert werden.
SMÜ SDL	In order to determine the starting line, can be marked with chalk this.
Nach-KS	Um die Startlinie festzustellen, kann mit Kreide **diese Linie** markiert werden.
SMÜ SDL	In order to determine the starting line, this line can be marked with chalk.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Daraufhin stiegen die Stilqualität um 1,25 Punkte und die Inhaltsqualität um 1,38 Punkte auf der Likert-Skala.

5.4.5.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.66 zeigt, waren die Inhalts- und Stilqualität in den meisten Fällen sowohl durchschnittlich als auch von ihren Intervallen vergleichbar.

Nur bei zwei MÜ-Systemen gab es signifikante Unterschiede (Tabelle 5.73): Bei dem NMÜ-System Google Translate sank die Stilqualität (− 5,1 %) (z (N = 24) = − 2,602 / p = ,009) und bei dem RBMÜ-System Lucy stieg die Inhaltsqualität (+ 4,7 %) (z (N = 14) = − 2,187 / p = ,029).

In Tabelle 5.74 fanden die Bewerter die Übersetzung von Google Translate stilistisch schlechter nach Verwendung des Bezugs anstatt des Pronomens (SQ:

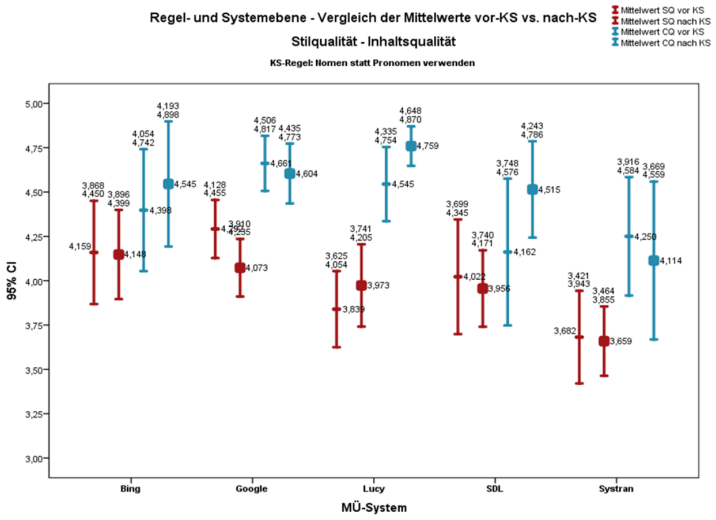


Abbildung 5.66: „Pronom. Bezüge verw." – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.73: „Pronom. Bezüge verw." – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	11	,685	− ,405	11	,210	− 1,254	11	,413	− ,819
Google	24	,009	− 2,602	24	,370	− ,897	24	,034	− 2,123
Lucy	14	,345	− ,944	14	,029	− 2,187	14	,089	− 1,699
SDL	17	,290	− 1,059	17	,247	− 1,157	17	,776	− ,284
Systran	11	,823	− ,223	11	,812	− ,238	11	,929	− ,089

p: Signifikanz z: Teststatistik nicht signifikant (p ≥ 0,05)

– ,50 Punkte auf der Likert-Skala), wobei die Bewertung der Inhaltsqualität im Durchschnitt unverändert blieb (CQ: 0,00 Punkte auf der Likert-Skala).

Tabelle 5.74: Beispiel 48

Vor-KS	Überprüfen Sie die Zuweisung des Ports im Geräte-Manager und stellen Sie **diesen** ggf. um.
GNMÜ	Check the assignment of the port in the Device Manager and, if necessary, modify it.
Nach-KS	Überprüfen Sie die Zuweisung des Ports im Geräte-Manager und stellen Sie **diesen Port** ggf. um.
GNMÜ	Check the assignment of the port in the Device Manager and, if necessary, modify this port.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Bei Lucy stieg z. B. im folgenden Satz (Tabelle 5.75) die Inhaltsqualität (CQ: + 0,75 Punkte auf der Likert-Skala) und die Stilqualität (SQ: + 0,38 Punkte auf der Likert-Skala).

Tabelle 5.75: Beispiel 49

Vor-KS	Ist der c-Faktor mit einer anderen Luftdichte angegeben worden, so ist **dieser** im Feld "Luftdichte" einzutragen.
RBMÜ Lucy	If the c-factor has been ***specified*** with a ***different*** air density, this ***must*** be entered in the "Air density" field.
Nach-KS	Ist der c-Faktor mit einer anderen Luftdichte angegeben worden, so ist **dieser Faktor** im Feld "Luftdichte" einzutragen.
RBMÜ Lucy	If the c-factor has been ***specified*** with a ***different*** air density, this factor ***must*** be entered in the "Air density" field.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.5.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwies sich nur bei dem HMÜ-System Systran eine signifikante starke negative Korrelation zwischen der Differenz in Fehlertyp SM.11 „Verwechslung des Sinns“ und der Differenz in der Stilqualität. (Tabelle 5.76)

Tabelle 5.76: „Pronom. Bezüge verw.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Systran		
	N	**p**	**ρ**
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl **SM.11 „Verwechs. des Sinns“**	11	,035	− ,638
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl **SM.11 „Verwechs. des Sinns“**	11	,071	− ,564
Differenz Q (nach KS − vor KS)			
Diff. der Anzahl **SM.11 „Verwechs. des Sinns“**	11	,038	− ,629

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Fehlertyp SM.11 erschien bei Systran sowohl vor als auch nach der Anwendung der KS-Regel (siehe Abbildung 5.62). Die negative Spearman-Korrelation besagt entsprechend, dass beim Auftreten dieses Fehlers, die Qualität deutlich sank. Umgekehrt stieg mit der Eliminierung dieses Fehlers die Qualität deutlich.

5.4.5.6 Vergleich der MÜ-Qualität vor vs. nach der Verwendung von pronominalen Bezügen auf Annotationsgruppenebene

Die Qualitätsunterschiede[39] waren nur bei den Annotationsgruppen FR und RF groß, während bei den Gruppen FF und RR das Qualitätsniveau relativ vergleichbar blieb (Abbildung 5.67).

[39] Definitionen der Qualität unter §4.5.5.1.

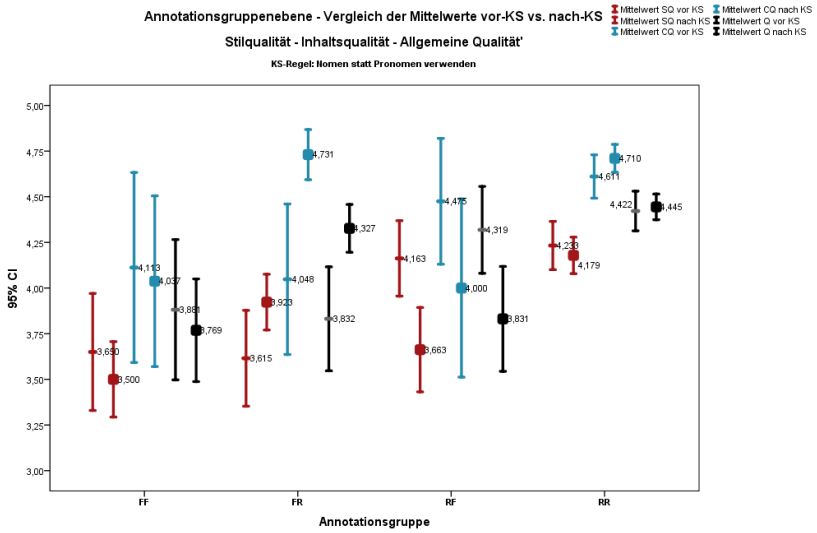


Abbildung 5.67: „Pronom. Bezüge verw." – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

In der Gruppe FF wurden die Sätze vor und nach der Anwendung der KS-Regel aus unterschiedlichen Gründen falsch übersetzt. Durchschnittlich sanken die Stil- und Inhaltsqualität leicht nachdem der Bezug wiederholt wurde (nach KS).

In der Gruppe FR stiegen erwartungsgemäß die Stil- und Inhaltsqualität signifikant (vgl. Tabelle 5.72), siehe Tabelle 5.77: bei der Stilqualität (z (N = 12) = − 2,547 / p = ,011) bzw. bei der Inhaltsqualität (z (N = 12) = − 3,072/ p = ,002). Wie die Aufteilung der Annotationsgruppen (siehe §5.4.5.3) zeigte, betrug der Anteil dieser Gruppe 18 % der analysierten Sätze bei dieser Regel. Die nach der Anwendung der Regel eliminierten Fehler waren unterschiedlich, sodass ein durch die Regel bestimmtes Fehlerbehandlungsmuster schwer erkennbar ist.

In der Gruppe RF sanken die Stil- und Inhaltsqualität signifikant (vgl. Tabelle 5.67), siehe Tabelle 5.77: bei der Stilqualität (z (N = 10) = − 2,509 / p = ,012) bzw. bei der Inhaltsqualität (z (N = 10) = − 2,313/ p = ,012). Der Anteil der Sätze in dieser Gruppe betrug 12 % der analysierten Sätze bei dieser Regel (siehe §5.4.5.3). Die nach der Regelanwendung vorgekommenen Fehler variierten.

Die Gruppe RR (Übersetzung vor und nach KS fehlerfrei) hatte den größten Anteil von 58 % der analysierten Sätze (siehe §5.4.5.3). Bei dieser Gruppe stieg durchschnittlich die Inhaltsqualität leicht, während bei der Stilqualität eine leichte Minderung verzeichnet wurde (Abbildung 5.67). Hierbei fanden die Bewerter, dass die wiederholte Erwähnung des Bezugs eine Präzisierung des Inhalts der

Tabelle 5.77: „Pronom. Bezüge verw." – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

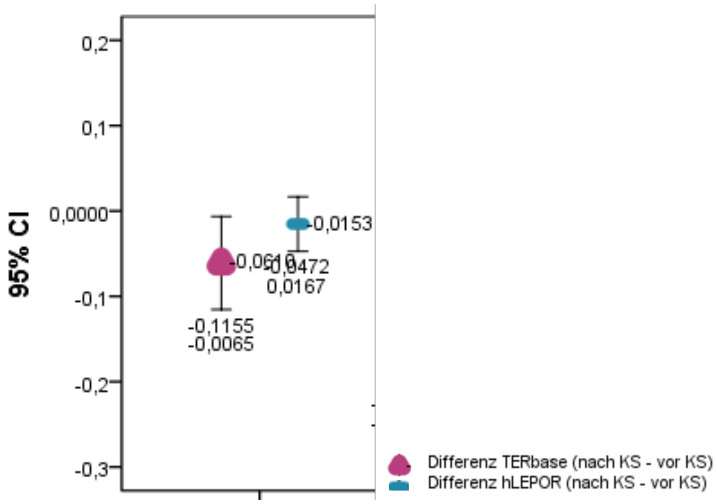
	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	8	,443	− ,768
Differenz CQ (nach KS − vor KS)	8	,799	− ,254
Differenz allg. Q (nach KS − vor KS)	8	,735	− ,338
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	12	,011	− 2,547
Differenz CQ (nach KS − vor KS)	12	,002	− 3,072
Differenz allg. Q (nach KS − vor KS)	12	,002	− 3,065
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	10	,012	− 2,509
Differenz CQ (nach KS − vor KS)	10	,012	− 2,313
Differenz allg. Q (nach KS − vor KS)	10	,012	− 2,527
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	47	,220	− 1,227
Differenz CQ (nach KS − vor KS)	47	,086	− 1,718
Differenz allg. Q (nach KS − vor KS)	47	,795	− ,260

MÜ ermöglichte, dennoch litt der Stil unter dieser Wiederholung. In Tabelle 5.70 war die MÜ vor und nach der Regelanwendung fehlerfrei. Nach der Regelanwendung stieg die Inhaltsqualität um 0,50, gleichzeitig sank die Stilqualität um 0,88 Punkte auf der Likert-Skala.

5.4.5.7 Vergleich der AEM-Scores vor vs. nach der Verwendung von pronominalen Bezügen sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores bei der Verwendung der Pronomen (vor KS) und nach der Wiederholung des Bezugs (nach KS) zeigte sowohl mit TERbase als auch

mit hLEPOR eine geringe nicht signifikante Verschlechterung der AEM-Scores (Abbildung 5.68).



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.68: „Pronom. Bezüge verw." – Mittelwert der Differenz der AEM-Scores

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei − ,061 (SD = ,240) und für die hLEPOR bei − ,015 (SD = ,141) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben) (Abbildung 5.68). Durch diese kleinen Unterschiede waren die Differenzen (nach KS − vor KS) in TERbase und hLEPOR nicht signifikant (z (N = 77) = − 1,962 / p = ,050) bzw. (z (N = 77) = − 1,221 / p = ,222). Eine mögliche Interpretation, warum die Differenz der AEM-Scores klein war: Bei dieser Regel gab es signifikante Veränderungen in der Fehleranzahl bei zwei Fehlertypen: der lexikalische Fehlertyp LX.6 „Konsistenzfehler" *stieg* (vlg. Tabelle 5.67) und der semantische Fehlertype SM.11 „Verwechslung des Sinns" *sank* (vlg. Tabelle 5.72) nach der Wiederholung des Bezugs (nach KS). In anderen Worten musste nach den AEMs in beiden Szenarien einer der beiden Fehler korrigiert werden, und zwar SM.11 vor KS und LX.6 nach KS. Dadurch näherten sich möglicherweise die AEM-Scores der beiden Szenarien an und fiel die Differenz zwischen den AEM-Scores klein aus.

5.4.5.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Durch das Auftreten von Fehlern sowohl vor als auch nach der Anwendung der KS konnte auch kein signifikanter Unterschied in der Qualität nachgewiesen werden (siehe §5.4.5.5). Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter mittlerer positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität. Tabelle 5.78 demonstriert die Korrelationswerte:

Tabelle 5.78: „Pronom. Bezüge verw." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	77	,003	,330	mittlerer Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	77	< ,001	,467	mittlerer Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Diese Korrelation zeigt, dass die Scores der beiden AEMs sich relativ synchron in die gleiche Richtung wie die Qualität bewegten.

5.4.5.8 Analyse der vierten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der vierten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (−)

steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[40]

Regel 4: Eindeutige pronominale Bezüge verwenden

Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Verwendung der pronominalen Bezüge

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene: H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. Die Fehleranzahl sank minimal nach der Verwendung der pronominalen Bezüge.	**Anz.F. (–)**
Auf Regel- und MÜ-Systemebene:	
Bei Google stieg die Fehleranzahl signifikant nach der Verwendung der pronominalen Bezüge.	**Go (+)**
Bei Bing blieb die Fehleranzahl unverändert.	**Bi (=)**
Bei den drei anderen Systemen sank die Fehleranzahl leicht.	**Lu (–)** **SD (–)** **Sy (–)**

[40] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Zweiter Analysefaktor

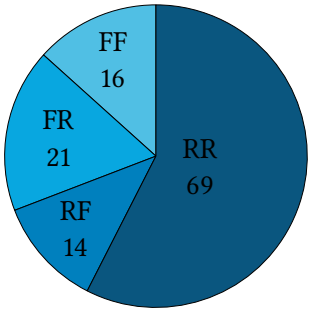


Abbildung 5.69: Aufteilung der Annotationsgruppen auf Regelebene

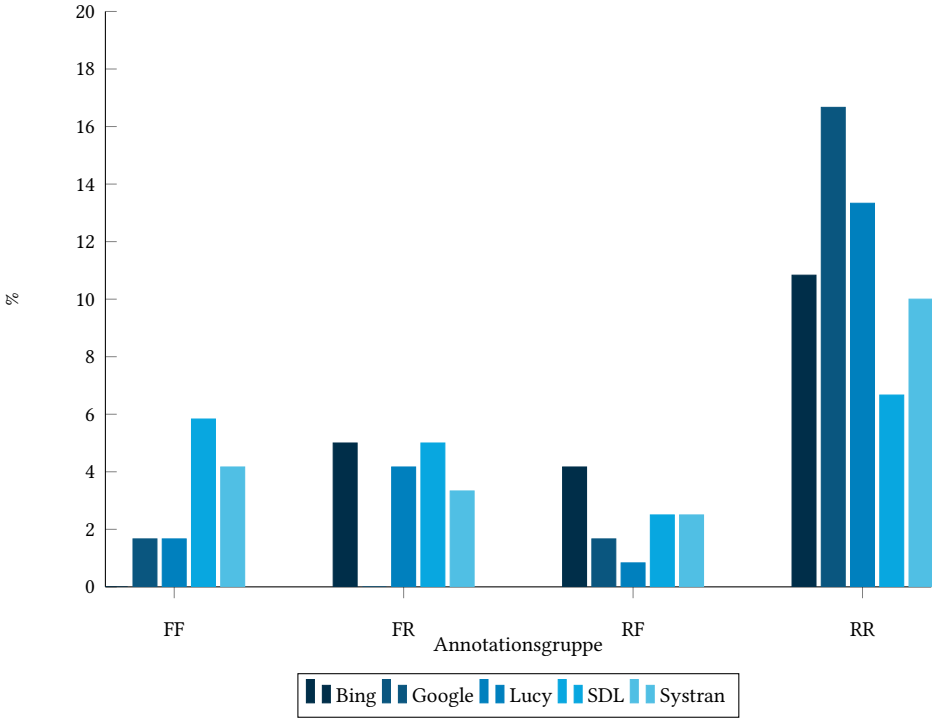


Abbildung 5.70: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Dritter Analysefaktor: Vergleich der Fehlertypen vor vs. nach der Verwendung der pronominalen Bezüge

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H1 wurde nur für zwei Fehlertypen bestätigt.
Die Fehleranzahl von LX.6 „Konsistenzfehler" stieg signifikant und die Fehleranzahl von SM.11 „Verwechslung des Sinns" sank signifikant nach der Verwendung der pronominalen Bezüge.

LX.6 (+)
SM.11 (–)

Auf Regel- und MÜ-Systemebene:
Auf Systemebene gab es keinen bestimmten Fehlertyp, der nach der Verwendung der pronominalen Bezüge statistisch signifikant beeinflusst wurde.
Dennoch wurde ein wesentlicher (nicht signifikanter) Anstieg in LX.6 „Konsistenzfehler" bei SDL und ein wesentlicher (nicht signifikanter) Rückgang in SM.11 „Verwechslung des Sinns" bei Lucy beobachtet.

SD:
LX.6 (+)

Lu:
SM.11 (–)

Alle weiteren Veränderungen der Fehlertypen waren sehr gering.

Vierter Analysefaktor: Vergleich der MÜ-Qualität vor vs. nach der Verwendung der pronominalen Bezüge

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. **SQ (–)** **CQ (+)**
Die Stilqualität sank leicht, während die Inhaltsqualität leicht stieg.

Auf Regel- und MÜ-Systemebene:
Die Stilqualität sank nur bei Google signifikant, während die Inhaltsqualität nur bei Lucy signifikant stieg. **SQ (–): Go** **CQ (+): Lu**

Die Qualitätsdifferenzen bei allen anderen Systemen fielen (sehr) gering aus und waren entsprechend nicht signifikant.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS – vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde nur für drei Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Falsche Wortstellung“ und SM.11 „Verwechslung des Sinns“ einzeln und der Differenz der Stilqualität.
Zudem bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.6 (Konsistenzfehler), GR.10 und SM.11 einzeln und der Differenz der Inhaltsqualität.

***neg* GR.10 <> SQ**
***neg* SM.11 <> SQ**

***neg* LX.6 <> CQ**
***neg* GR.10 <> CQ**
***neg* SM.11 <> CQ**

Auf Regel- und MÜ-Systemebene:
Die einzige signifikante Korrelation bestand bei Systran und es handelte sich dabei um einen signifikanten starken negativen Zusammenhang zwischen der Differenz der Fehleranzahl des SM.11 „Verwechslung des Sinns“ und der Differenz der Stilqualität.

Sy
***neg* SM.11 <<>> SQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität vor vs. nach der Verwendung der pronominalen Bezüge auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde für die Gruppen FR und RF bestätigt:
Bei der Annotationsgruppe FR stiegen die Stil- und Inhaltsqualität signifikant nach der Verwendung der pronominalen Bezüge.

SQ (+)
CQ (+)

Bei der Annotationsgruppe RF sanken die Stil- und Inhaltsqualität signifikant.

SQ (–)
CQ (–)

Bei der Annotationsgruppe FF gingen die Stil- und Inhaltsqualität leicht zurück.

SQ (–)
CQ (–)

Bei der Annotationsgruppe RR sank die Stilqualität leicht und die Inhaltsqualität nahm leicht zu.

SQ (–)
CQ (+)

Siebter Analysefaktor: Vergleich der AEM-Scores vor vs. nach der Verwendung der pronominalen Bezüge

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden.
Die AEM-Scores von TERbase und hLEPOR verschlechtern sich nur leicht nach der Verwendung der pronominalen Bezüge.

TERbase (–)
hLEPOR (–)

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen den Differenzen der Scores der beiden AEMs (TERbase und hLEPOR) und der Differenz der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.6 FÜNFTE REGEL: Partizipialkonstruktionen vermeiden

5.4.6.1 Überblick

Im Folgenden wird die KS-Regel „Partizipialkonstruktionen vermeiden“ kurz beschrieben.[41] Zudem wird zusammenfassend und anhand eines Beispiels die Um-

[41] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

setzung der Regel bei der Analyse demonstriert. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Partizipialkonstruktionen vermeiden (tekom-Regel-Nr. S 303)

Nach dieser Regel (tekom 2013: 70) soll statt der Partizipialkonstruktion eine einfache Satzstruktur mit mehreren kurzen Sätzen oder Nebensätzen verwendet werden.

Begründung: Die Partizipialkonstruktion erschwert die Textverständlichkeit (ebd.).

Umsetzungsmuster:

Vor KS: Der Satz beinhaltet eine Partizipialkonstruktion.
Nach KS: aus der Partizipialkonstruktion einen Nebensatz bauen

KS-Stelle

Vor KS: alle Wörter, die das Nomen beschreiben, angefangen mit dem Artikel, falls vorhanden.
Nach KS: die konvertierten Wörter der Version vor KS (inkl. des Kommas innerhalb der KS-Stelle)

Beispiele

Speziell auf diese Lautsprecher abgestimmtes Zubehör erhalten Sie in unserem Webshop.

Zubehör, das speziell auf diese Lautsprecher abgestimmt ist , erhalten Sie in unserem Webshop.

Aufteilung der Testsätze: Der Datensatz besteht aus 24 verschiedenen Partizipialkonstruktionen, die unterschiedliche Längen haben und an unterschiedlichen Stellen in den Sätzen erscheinen.

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.6.2 Vergleich der Fehleranzahl mit und ohne die Verwendung von Partizipialkonstruktionen

Die Fehleranzahl stieg um 26,4 % von 110 Fehlern im Falle der Verwendung von Partizipialkonstruktionen (M = ,92 / SD = 1,089 / N = 120) auf 139 Fehler bei der

Vermeidung der Partizipialkonstruktionen (M = 1,16 / SD = 1,167 / N = 120) (Abbildung 5.71 und Abbildung 5.72). Der Mittelwert der Differenz (nach KS − vor KS) der Fehleranzahl pro Satz lag somit bei ,24 (SD = 1,277) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,01 (SD = 1,087) und einem Maximum von ,47 (SD = 1,464) (Bootstrapping mit 1000 Stichproben). Die Differenz (nach KS − vor KS) der Fehleranzahl erwies sich als signifikant (z (N = 120) = − 1,980 / p = ,048).

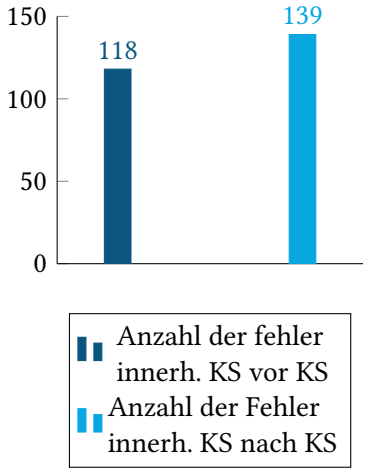


Abbildung 5.71: „Partizipialkonst. verm." – Fehlersumme vor vs. nach KS

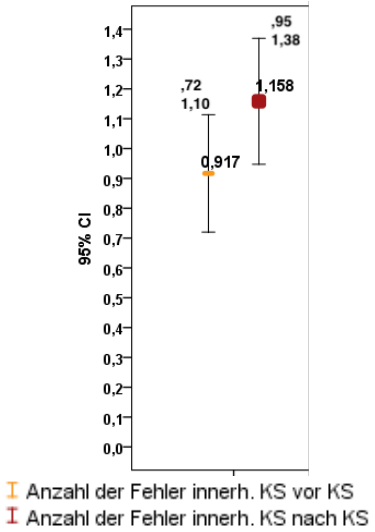


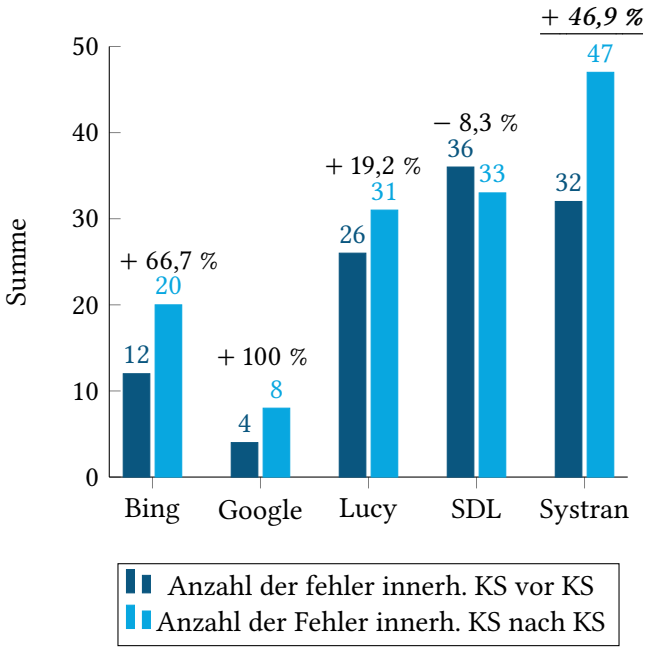
Abbildung 5.72: „Partizipialkonst. verm." – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Eine genauere Untersuchung der Fehlertypen anhand von Beispielen ist unter §5.4.6.4.

5.4.6.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Wie Abbildung 5.73 zeigt, stieg die Fehleranzahl bei allen Systemen außer dem SMÜ-System SDL nachdem die Sätze ohne Partizipialkonstruktionen formuliert wurden, allerdings war der Rückgang bei SDL insignifikant (Mdiff = − 0,125).

Die einzige signifikante Veränderung nach der Anwendung der KS-Regel entstand bei dem HMÜ-System Systran Mdiff = ,625; (z (N = 24) = − 2,156 / p = ,031). Die weiteren Anstiege waren nicht signifikant: Google Mdiff = ,167; Bing Mdiff = ,333; Lucy Mdiff = ,208.



Signifikante Differenz vor vs. nach KS

Abbildung 5.73: „Partizipialkonst. verm." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Wie Abbildung 5.73 zeigt, verzeichnete Google die niedrigste Fehleranzahl sowohl vor als auch nach der Regelanwendung; gleichzeitig stieg die Fehleranzahl nach KS. Auf der anderen Seite konnte das SMÜ-System SDL in manchen Fällen von der Regelanwendung profitieren, wie Tabelle 5.79 demonstriert.

Nachdem die Partizipialkonstruktion vermieden wurde, konnten der Wortstellungsfehler (GR10) und Auslassungsfehler (LX.3) in der KS-Stelle (in ‚For the machine tools required') korrigiert werden, während das NMÜ-System Google Translate in der Lage war, den Satz mit und ohne Partizipialkonstruktion fehlerfrei zu übersetzen.

5.4.6.3 Aufteilung der Annotationsgruppen

Wie Abbildung 5.74 zeigt, war die größte Annotationsgruppe bei dieser Regel die Gruppe FF; 42 % der Sätze wurden mit und ohne Partizipialkonstruktion falsch übersetzt. Unter §5.4.6.4 werden die persistenten Fehlertypen ins Visier genommen. Die zweitgrößte Gruppe war RR, repräsentiert mit 28 %. Das zeigt, dass die

Tabelle 5.79: Beispiel 50

Vor-KS	**Die für die Maschine benötigten Werkzeuge** sind im Lieferumfang nicht enthalten.
SMÜ SDL	For the machine XXX tools required are not included in the delivery.
GNMÜ	The tools required for the machine are not included in the delivery.
Nach-KS	**Die Werkzeuge, die für die Maschine benötigt werden**, sind im Lieferumfang nicht enthalten.
SMÜ SDL	The tools that are required for the machine are not included in the delivery.
GNMÜ	The tools required for the machine are not included in the delivery.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Systeme mehr als ein Viertel der Sätze mit und ohne Partizipialkonstruktion fehlerfrei übersetzen konnten. Im nächsten Abschnitt werden die einzelnen Systeme auf Annotationsgruppenebene im Detail untersucht.

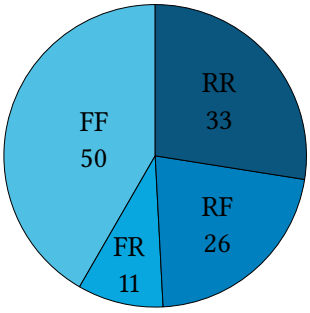


Abbildung 5.74: „Partizipialkonst. verm." – Aufteilung der Annotationsgruppen

Bei 22 % der Sätze traten Fehler erst nach der Anwendung der Regel (Gruppe RF) auf, siehe Abbildung 5.74. Dies ist der höchste Prozentsatz der Gruppe RF unter allen analysierten Regeln. Wie oben erwähnt, war die Anwendung der Regel oft mit einer falschen Kommaplatzierung bei der Verwendung von Relativpronomen (‚which' vs. ‚that') verbunden, siehe Tabelle 5.80. Schließlich beinhaltete die

MÜ in 9 % der Fälle mindestens einen Fehler vor der Regelanwendung, der nach der Regelanwendung behoben wurde (Gruppe FR).

5.4.6.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Mehr als die Hälfte der Sätze wurden bei drei MÜ-Systemen, dem RBMÜ-System Lucy (67 %), SMÜ-System SDL (54 %) und HMÜ-System Systran (54 %), mit und ohne Partizipialkonstruktion (d. h. sowohl vor als auch nach der Regelanwendung) falsch übersetzt (Annotationsgruppe FF). Bei den beiden HMÜ-Systemen Bing und Systran wurde jeweils ein Drittel der Übersetzungen (33 %) mit Partizipialkonstruktion (vor KS) richtig und ohne Partizipialkonstruktion (nach KS) falsch übersetzt (Annotationsgruppe RF), mehr zu den Fehlertypen unter §5.4.6.4. Somit hatten die verschiedenen älteren MÜ-Ansätze Schwierigkeiten mit der Übersetzung in den beiden Szenarien.

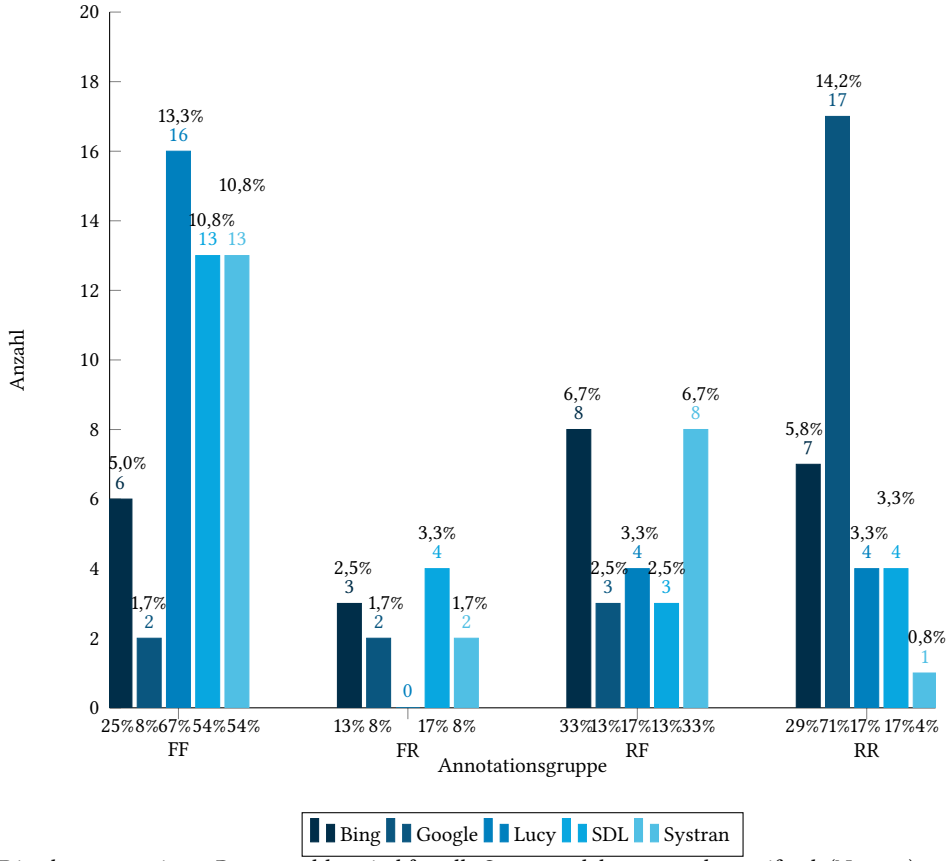
Dies steht in Kontrast zum neuen Ansatz der NMÜ von Google, bei dem 71 % der Übersetzungen sowohl vor als auch nach der Anwendung der KS-Regel fehlerfrei waren (Annotationsgruppe RR).

5.4.6.4 Vergleich der Fehlertypen mit und ohne die Verwendung von Partizipialkonstruktionen

Nach der Anwendung dieser Regel verändert sich die Anzahl zweier Fehlertypen deutlich, und zwar stieg die Fehleranzahl bei Fehlertyp OR.1 „Orthografie – Zeichensetzung“ und sank bei Fehlertyp GR.10 „Grammatik – Falsche Wortstellung“. Beide Veränderungen erwiesen sich als signifikant $p < ,001$ beim Fehlertyp OR.1 bzw. $p = ,011$ beim Fehlertyp GR.10 ($N = 120$).

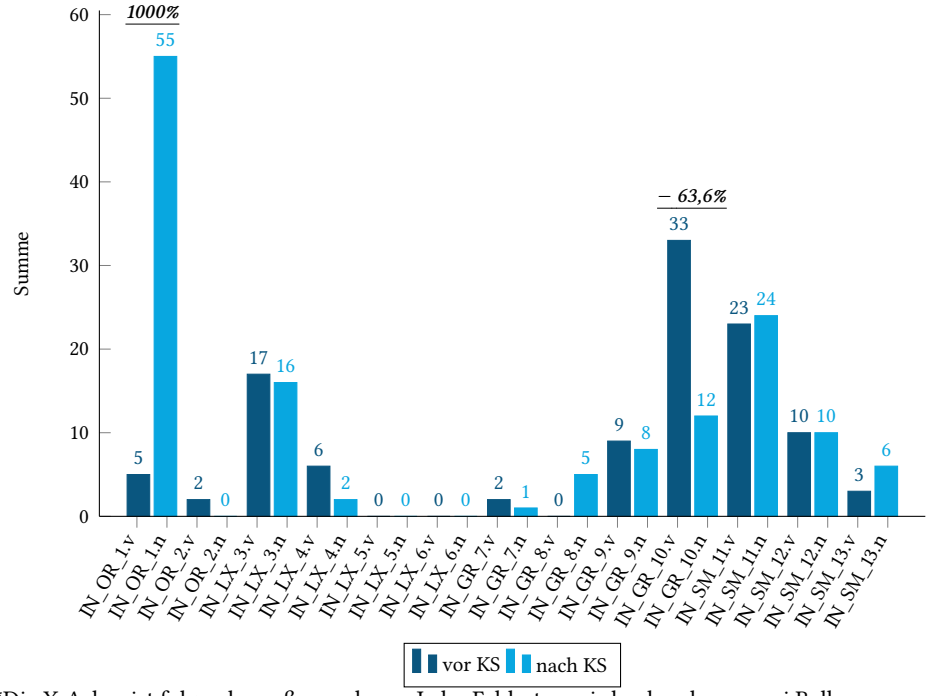
Um diese KS-Regel umzusetzen, wird die Partizipialkonstruktion aufgelöst, indem daraus einen Nebensatz gebildet wird. Die Übersetzung des Nebensatzes ins Englische erfordert eine Ermittlung des adäquaten Relativpronomens, d. h. eine Unterscheidung zwischen Relativpronomen wie ‚which‘ und ‚that‘. Dieser Unterscheidung bedarf in der Regel Kontextinformationen und nicht selten ist sie auch für den Humanübersetzer problematisch (vgl. Swan 1980: 527ff.).[42] Da die Entscheidung über Relativpronomen wie ‚which‘ und ‚that‘ engverbunden mit

[42] Man unterscheidet im Englischen zwischen restriktiven und nicht-restriktiven Relativsätzen: Im restriktiven Relativsatz wird die Bedeutung des Nomens, das beschrieben wird, begrenzt. Ohne den restriktiven Relativsatz ändert sich die Bedeutung des gesamten Satzes. Für restriktive Relativsätze verwendet man ‚that‘. Der nicht-restriktiver Relativsatz bietet lediglich zusätzliche Informationen über das Nomen und kann ohne Einfluss auf die Bedeutung entfernt werden. Für nicht-restriktive Relativsätze verwendet man ‚which‘. (vgl. Swan 1980: 527ff.)



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.75: „Partizipialkonst. verm.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler vor KS und der zweite die Summe der Fehler nach KS, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
**__*Signifikante Differenz vor vs. nach KS*__
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.76: „Partizipialkonst. verm.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

der Verwendung bzw. Nichtverwendung von Kommas ist (ebd.), stieg die Frequenz des Fehlertyps OR.1 „Orthografie – Zeichensetzung“ jedes Mal, wenn ein falsches Relativpronomen bei der Übersetzung verwendet wurde, und zwar von 5 auf 55 (1000 % / Mv = ,04 / SDv = ,239 / Mn = ,46 / SDn = ,660 / N = 120). In Tabelle 5.80 fehlte nach der Regelanwendung das Komma bevor ‚which‘.

Tabelle 5.80: Beispiel 51

Vor-KS	Das Gerät verbindet sich mit **der neu gewählten Netzwerkadresse.**
GNMÜ	The device connects to the newly selected network address.
Nach-KS	Das Gerät verbindet sich mit **der Netzwerkadresse, die neu gewählt wird.**
GNMÜ	The device connects to the network address, which is selected again.

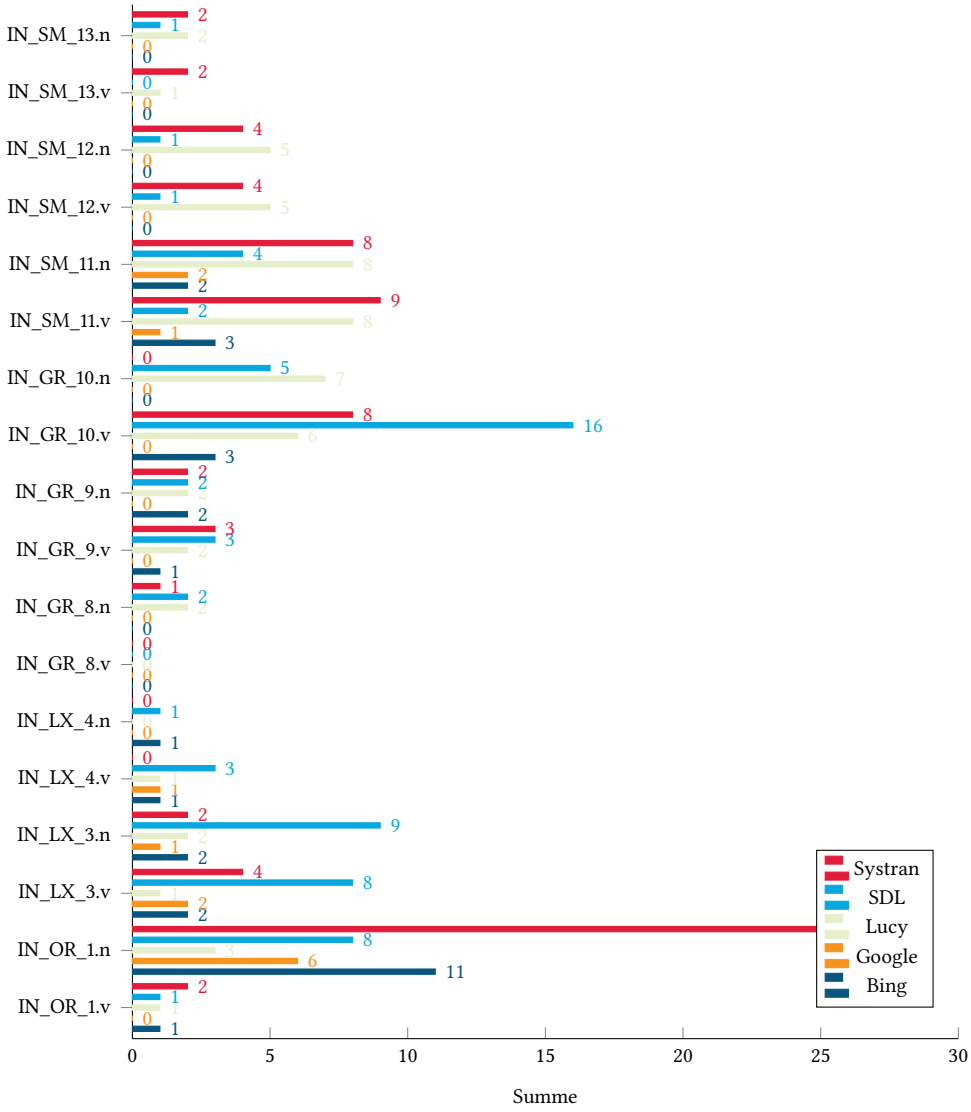
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Auf der anderen Seite erschwerten Partizipialkonstruktionen, insbesondere lange Partizipialkonstruktionen, den MÜ-Systemen die Aufgabe, eine korrekte Wortstellung zu produzieren. Eine Auflösung der Partizipialkonstruktion (nach KS) konnte helfen, die Fehleranzahl vom Fehlertyp GR.10 „Grammatik – Falsche Wortstellung“ von 33 auf 12 (– 64 % / Mv = ,28 / SDv = ,579 / Mn = ,10 / SDn = ,301 / N = 120) zu reduzieren (siehe Tabelle 5.82).

5.4.6.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt, dass der allgemein signifikante Unterschied in der Fehleranzahl des Fehlertyps OR.1 „Orthografie – Zeichensetzung“ bei den Systemen Bing, SDL und Systran und des Fehlertyps GR.10 „Grammatik – Falsche Wortstellung“ nur bei dem System SDL zu beobachten war (Abbildung 5.77). Bei den Systemen Google und Lucy gab es keinen bestimmten Fehlertyp, dessen Fehleranzahl nach der Regelanwendung eine signifikante Änderung aufwies. Darüber hinaus fiel die Fehleranzahl aller Fehlertypen bei Google sowohl vor als auch nach der Regelanwendung sehr gering aus.

Der Fehlertyp OR.1 stieg deutlich bei Bing von 1 auf 11 (+ 1000 %), bei SDL von 1 auf 8 (+ 700 %) und bei Systran von 2 auf 28 (+ 1300 %). Der Fehlertyp GR.10



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 5 und 6 bei gar keinem MÜ-System vor. Zudem kamen die Fehlertypen 2 und 7 nur einmal jeweils bei 2–3 MÜ-Systemen in vereinzelten Fällen vor.

Abbildung 5.77: „Partizipialkonst. verm.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

sank bei SDL von 16 auf 5 (− 68,8 %). Die Differenz in der Fehleranzahl der beiden Typen OR.1 und GR.10 erwies sich in den genannten Systemen folgendermaßen als signifikant (Tabelle 5.81).

Tabelle 5.81: „Partizipialkonst. verm." – Fehlertypen mit signifikanter Veränderung nach KS

	N	Mittelwert	Standard-abweichung	Signifikanz (McNemar-Test)
OR.1 „Zeichensetzung"				
Bing	24	vor KS = ,04	vor KS = ,204	p = ,002
		nach KS = ,46	nach KS = ,509	
SDL	24	vor KS = ,04	vor KS = ,204	p = ,016
		nach KS = ,33	nach KS = ,482	
Systran	24	vor KS = ,08	vor KS = ,408	p < ,001
		nach KS = 1,17	nach KS = ,702	
GR.10 „Falsche Wortstellung"				
SDL	24	vor KS = ,67	vor KS = ,761	p = ,016
		nach KS = ,21	nach KS = ,415	
Systran	24	vor KS = ,33	vor KS = ,702	p = ,037
		nach KS = ,38	nach KS = ,576	

Bei dem SMÜ-System SDL und dem HMÜ-System Systran wurde der Fehlertyp GR.10 „Grammatik – Falsche Wortstellung" nach dem Zerlegen der Partizipialkonstruktion (nach KS) behoben. Dieses Ergebnis ist in SMÜ-Systemen im Falle von Partizipialkonstruktionen, die in den Trainingsdaten nicht (häufig) vorkommen, zu erwarten, da das Zerlegen solcher Partizipialkonstruktionen das SMÜ-System dabei unterstützt, eine korrekte Wortstellung zu produzieren.

Tabelle 5.82 zeigt, wie das Vermeiden der Partizipialkonstruktion SDL u. a. dabei unterstützte, den Wortstellungsfehler zu korrigieren.

5.4.6.5 Vergleich der MÜ-Qualität mit und ohne die Verwendung von Partizipialkonstruktionen sowie die Korrelation zwischen den Fehlertypen und der Qualität

Nachdem die Partizipialkonstruktionen vermieden wurden (nach KS), sanken sowohl die Stil- als auch die Inhaltsqualität.[43] Auf den ersten Blick erkennt man in

[43] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.82: Beispiel 52

Vor-KS	Durch Eingabe **der mit einem roten Sternchen gekennzeichneten Parameter** erfolgt die minimale Konfigurierung.
SMÜ SDL	By entering the marked with a red asterisk parameter, the minimum configuration is performed.
Nach-KS	Durch Eingabe **der Parameter, die mit einem roten Sternchen gekennzeichnet sind**, erfolgt die minimale Konfigurierung.
SMÜ SDL	By entering the parameters that are marked with a red asterisk, the minimum configuration is performed.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Abbildung 5.79, dass der Einfluss auf die Stilqualität im Vergleich zur Inhaltsqualität größer ausfiel. Die Stilqualität sank um 7,8 % (Mv = 3,97 / SDv = ,658 / Mn = 3,66 / SDn = ,535 / N = 98). Die Inhaltsqualität sank um 1,7 % (Mv = 4,23 / SDv = ,789 / Mn = 4,16 / SDn = ,726 / N = 98) (Abbildung 5.78). Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei − ,312 (SD = ,575) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,428 und einem Maximum von − ,196 und für die Inhaltsqualität bei − ,066 (SD = ,714) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,210 und einem Maximum von ,078 (Bootstrapping mit 1000 Stichproben) (Abbildung 5.79). Nur die Differenz (nach KS − vor KS) in der Stilqualität erwies sich als hochsignifikant (z (N = 98) = − 4,997 / p < ,001). Bei der Inhaltsqualität war die Differenz insignifikant (z (N = 98) = − ,597 / p = ,550).

Tabelle 5.83 zeigt einen Satz, der zwar mit und ohne Partizipialkonstruktion fehlerfrei übersetzt wurde, allerdings sanken sowohl die Stilqualität (− ,75 Punkte auf der Likert-Skala) als auch die Inhaltsqualität (− ,25 Punkte auf der Likert-Skala), nachdem die Partizipialkonstruktion vermieden wurde (nach KS).

Eine genaue Betrachtung der in der Humanevaluation identifizierten Qualitätskriterien (Abbildung 5.80) verrät, dass der Rückgang bei der Stilqualität überwiegend an der unklaren Darstellung (SQ1) sowie der unnatürlichen Formulierung (SQ3) und bei der Inhaltsqualität insbesondere an der beeinträchtigen Verständlichkeit (CQ2) lag.

In Tabelle 5.83 fanden die Bewerter die MÜ vor der Anwendung der KS-Regel (mit Partizipialkonstruktion) prägnanter und idiomatischer. In einem weiteren Beispiel, in dem die MÜ mit und ohne Partizipialkonstruktion Fehler beinhaltete,

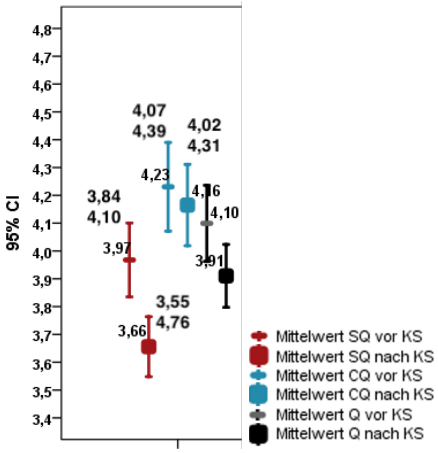


Abbildung 5.78: „Partizipialkonst. verm.“ – Mittelwerte der Qualität vor und nach KS

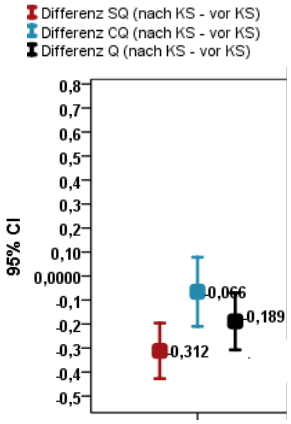
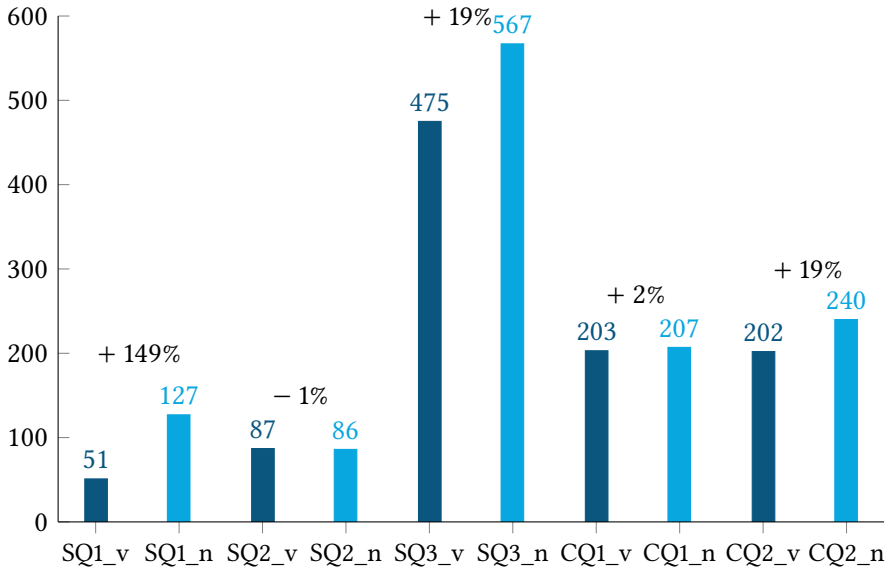


Abbildung 5.79: „Partizipialkonst. verm.“ – Mittelwert der Qualitätsdifferenzen

Tabelle 5.83: Beispiel 53

Vor-KS	**Die für Ihren Flug erlaubte Freigepäckmenge** ist auf Ihrem Flugschein angegeben.
HMÜ Systran	The free luggage quantity permitted for your flight is provided on your ticket.
Nach-KS	**Die Freigepäckmenge, die für Ihren Flug erlaubt ist**, ist auf Ihrem Flugschein angegeben.
HMÜ Systran	The free luggage quantity, which is permitted for your flight, is provided on your ticket.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder. **CQ2:** Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.80: „Partizipialkonst. verm.“ – Vergleich der Qualitätskriterien

sanken die Stilqualität (− ,63 Punkte auf der Likert-Skala) und die Inhaltsqualität (− ,50 Punkte auf der Likert-Skala) nach der Regelanwendung (Tabelle 5.84).

Die Wortstellungsfehler (GR.10) wurden nach der Regelanwendung nicht behoben. Zudem kamen zwei neue Fehler hinzu: der orthografische Fehler OR.1 (fehlendes Komma) und der lexikalische Fehler LX.3 (Auslassen des Relativpronomens ‚that‘). Entsprechend wurden die Verständlichkeit und Idiomatik der MÜ beeinträchtigt.

5.4.6.6 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung bei der Fehleranzahl jedes Fehlertyps (Anz. nach KS − Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS − Q. vor KS) zusammenhängt. Mithilfe des Spearman-Tests

Tabelle 5.84: Beispiel 54

Vor-KS	**Die für Ihren Flug erlaubte Freigepäckmenge** ist auf Ihrem Flugschein angegeben.
SMÜ SDL	The for your flight allowed free baggage allowance is provided on your ticket.
Nach-KS	**Die Freigepäckmenge, die für Ihren Flug erlaubt ist**, ist auf Ihrem Flugschein angegeben.
SMÜ SDL	The free baggage allowanceXXX for your flight XXX is allowed, is provided on your ticket.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz im Fehlertyp LX.4 und der Differenz in der Stilqualität. Außerdem erwies sich ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz in den Fehlertypen OR.1, LX.3, GR.9, GR.10 und SM.11 einzeln und der Differenz in der Stilqualität. (Tabelle 5.85)

Bezüglich der Inhaltsqualität erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.3, LX.4 und GR.10 einzeln und der Differenz in der Inhaltsqualität sowie ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz in den Fehlertypen GR.9 und SM.11 einzeln und der Differenz in der Inhaltsqualität. (Tabelle 5.85)

Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht erwiesen werden.

Diese signifikanten negativen Korrelationen deuten darauf hin, dass sobald die Fehleranzahl der genannten Fehlertypen sank, die Qualität zunahm, und umgekehrt. In Tabelle 5.86 wurden der lexikalische Fehler LX.4 „Zusätzliches Wort eingefügt“ (in ‚coordinated‘) und der Wortstellungsfehler GR.10 (in ‚accessories‘) eliminiert, daraufhin stieg die Qualität deutlich.

Die Stilqualität verbesserte sich um 1,25 Punkte und die Inhaltsqualität um 1,75 Punkte auf der Likert-Skala nach der Formulierung des Satzes ohne Partizipialkonstruktion (nach KS).

Tabelle 5.85: „Partizipialkonst. verm.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl der **OR.1 „Zeichensetzung“**	97	,006	− ,278
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	97	,004	− ,293
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	97	,001	− ,334
Diff. der Anzahl der **GR.9 „Kongruenzfehler“**	97	,038	− ,211
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	97	,010	− ,260
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	97	,009	− ,265
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl der **OR.1 „Zeichensetzung“**	97	,395	− ,087
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	97	< ,001	− ,462
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	97	,001	− ,338
Diff. der Anzahl der **GR.9 „Kongruenzfehler“**	97	,014	− ,250
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	97	,002	− ,307
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	97	,023	− ,230
Differenz allg. Q (nach KS − vor KS)			
Diff. der Anzahl der **OR.1 „Zeichensetzung“**	97	,037	− ,212
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	97	< ,001	− ,413
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	97	< ,001	− ,350
Diff. der Anzahl der **GR.9 „Kongruenzfehler“**	97	,011	− ,257
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	97	,001	− ,324
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	97	,010	− ,260

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Tabelle 5.86: Beispiel 55

Vor-KS	**Speziell auf diese Lautsprecher abgestimmtes Zubehör** erhalten Sie in unserem Webshop.
SMÜ SDL	Designed specifically for these speakers coordinated accessories are available in our webshop.
Nach-KS	**Zubehör, das speziell auf diese Lautsprecher abgestimmt ist**, erhalten Sie in unserem Webshop.
SMÜ SDL	Accessories that are specially adapted to these speakers are available in our webshop.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.6.6.1 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.81 zeigt, sank die Stilqualität signifikant bei dem HMÜ-System Bing (SQ − 13,9 %), dem MNÜ-System Google Translate (SQ − 9,4 %) sowie bei dem RBMÜ-System Lucy (SQ − 6,2 %). Bei dem SMÜ-System SDL und dem HMÜ-System Systran sank ebenfalls die Stilqualität nach der Regelanwendung, allerdings war der Rückgang nicht groß und entsprechend insignifikant.

Auf der anderen Seite blieb die Inhaltsqualität vor und nach der Regelanwendung bei allen Systemen bis auf das HMÜ-System Systran ohne große Veränderung: Bei dem HMÜ-System Systran stieg die Inhaltsqualität signifikant um 6,2 %. Gleichzeitig sank die Inhaltsqualität bei dem anderen HMÜ-System Bing mit einem insignifikanten Prozentsatz von − 7,9 %. Bei den anderen Systemen war die Veränderung minimal und entsprechend insignifikant.

Tabelle 5.87 zeigt, dass der Rückgang in der Stilqualität bei drei Systemen, nämlich Bing, Google und Lucy, signifikant war. Bei Bing und Google waren die größten Annotationsgruppen RR und RF, d. h. dass die Mehrheit der Sätze vor der Regelanwendung fehlerfrei übersetzt wurde (siehe §5.4.6.3). Eine korrekt übersetzte Partizipialkonstruktion zeigte sich in der Analyse der Qualitätskriterien als idiomatischer bzw. stilistischer als eine korrekt übersetzte aufgelöste Partizipialkonstruktion. Dies erklärt den signifikanten Rückgang der Stilqualität bei Bing und Google.

Bei Lucy wurde die Mehrheit der Sätze vor und nach der Regelanwendung falsch übersetzt (Gruppe FF 67 %, siehe Abbildung 5.75). Die Humanevaluation zeigte, dass knapp 65 % (9 von 14 Sätzen) der Sätze der FF Gruppe bei Lucy nach

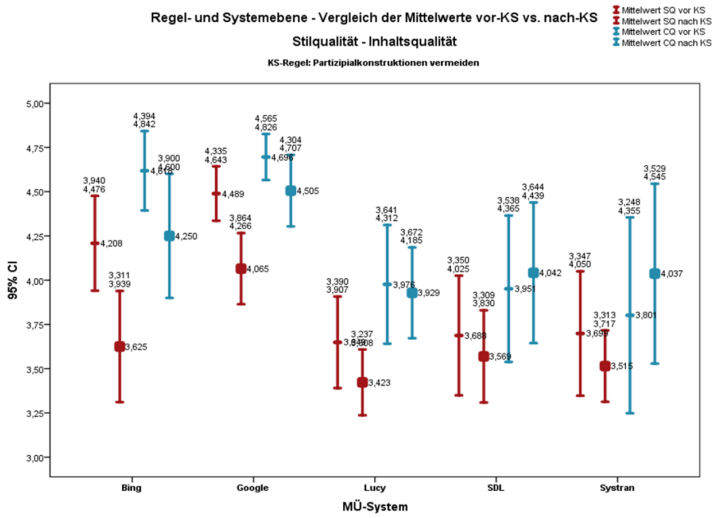


Abbildung 5.81: „Partizipialkonst. verm." – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.87: „Partizipialkonst. verm." – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	18	,003	− 3,004	18	,088	− 1,707	18	,009	− 2,616
Google	23	,001	− 3,261	23	,094	− 1,674	23	,001	− 3,436
Lucy	21	,006	− 2,734	21	,830	− ,214	21	,058	− 1,896
SDL	18	,522	− ,641	18	,711	− ,371	18	,868	− ,166
Systran	17	,124	− 1,538	17	,039	− 2,064	17	,812	− ,238

p: Signifikanz z: Teststatistik nicht signifikant (p ≥ 0,05)

der Regelanwendung stilistisch schlechter bewertet wurden. In Tabelle 5.88 sank die Inhaltsqualität (– ,63 Punkte auf der Likert-Skala), während die Inhaltsqualität anstieg (+ 0,38 Punkte auf der Likert-Skala) – nach der Regelanwendung.

Tabelle 5.88: Beispiel 56

Vor-KS	**Die in den Bedienungsanweisungen der eingebauten Geräte vorgeschriebenen Gebrauchsbedingungen** müssen strikt eingehalten werden.
RBMÜ Lucy	The use conditions stipulated in the service instructions of the built-in devices must be strictly adhered to.
Nach-KS	**Die Gebrauchsbedingungen, die in den Bedienungsanweisungen der eingebauten Geräte vorgeschrieben sind**, müssen strikt eingehalten werden.
RBMÜ Lucy	The use conditions which are stipulated in the service instructions of the built-in devices must be strictly adhered to.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

5.4.6.6.2 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwies sich bei dem HMÜ-System Bing ein stark signifikanter negativer Zusammenhang zwischen Fehlertyp GR.10 „Wortstellungsfehler“ und der Stilqualität sowie ein signifikanter mittlerer negativer Zusammenhang zwischen Fehlertyp SM.11 „Verwechslung des Sinns“ und der Inhaltsqualität (Tabelle 5.89). Bei dem HMÜ-System Systran erwies sich nur eine signifikante starke negative Korrelation zwischen Fehlertyp LX.3 „Wort ausgelassen“ und der Inhaltsqualität (Tabelle 5.89).

Bei dem SMÜ-System SDL gab es schließlich mehrere signifikante Korrelationen: Bei der Stilqualität erwies sich ein signifikanter starker negativer Zusammenhang zwischen Fehlertyp LX.3, LX.4 und GR.10 einzeln und der Stilqualität. Bei der Inhaltsqualität erwies sich ein signifikanter starker negativer Zusammenhang zwischen Fehlertyp OR.1, LX.3, LX.4 und GR.10 einzeln und der Inhaltsqualität (siehe Tabelle 5.89 und Tabelle 5.86).

Tabelle 5.89: „Partizipialkonst. verm.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Bing			SDL			Systran		
	N	p	ρ	N	p	ρ	N	p	ρ
Differenz der Anzahl SQ (nach KS − vor KS)									
OR.1 „Zeichen“				18	,052	− ,464			
LX.3 „W. fehlt“				18	,002	− ,684	17	,275	− ,281
LX.4 „W. extra“				18	,011	− ,586			
GR.9 „Kongru.“				18	,076	− ,429			
GR.10 „Wortst.“	18	,028	− ,518	18	,008	− ,604			
SM.11 „Sinn“	18	,091	− ,410						
Differenz der Anzahl CQ (nach KS − vor KS)									
OR.1 „Zeichen“				18	,003	− ,656			
LX.3 „W. fehlt“				18	< ,001	− ,754	17	,024	− ,545
LX.4 „W. extra“				18	,017	− ,555			
GR.9 „Kongru.“				18	,080	− ,424			
GR.10 „Wortst.“	18	,765	− ,076	18	,005	− ,629			
SM.11 „Sinn“	18	,039	− ,490						
Differenz der Anzahl Q (nach KS − vor KS)									
OR.1 „Zeichen“				18	,007	− ,608			
LX.3 „W. fehlt“				18	< ,001	− ,758	17	,047	− ,487
LX.4 „W. extra“				18	,009	− ,595			
GR.9 „Kongru.“				18	,039	− ,490			
GR.10 „Wortst.“	18	,160	− ,346	18	,002	− ,689			
SM.11 „Sinn“	18	,011	− ,581						

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

5.4.6.7 Vergleich der MÜ-Qualität mit und ohne die Verwendung von Partizipialkonstruktionen auf Annotationsgruppenebene

Die Qualitätsveränderung[44] der MÜ variierte nach Vermeidung der Partizipialkonstruktion (nach KS) in den verschiedenen Annotationsgruppen (Abbildung 5.82).

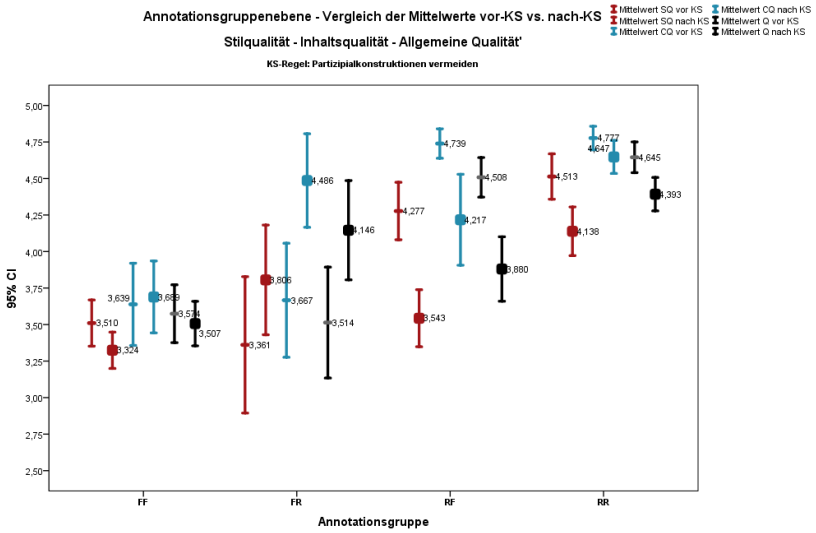


Abbildung 5.82: „Partizipialkonst. verm." – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

Die Gruppe FF (Übersetzung vor und nach KS falsch) hatte den größten Anteil von 42 % der analysierten Sätze (siehe §5.4.6.3). In dieser Gruppe stieg zwar die Inhaltsqualität leicht, allerdings sank die Stilqualität signifikant (N = 37) = − 2,623 / p = ,008), siehe Tabelle 5.91. Die Bewerter fanden in vielen Fällen die MÜ bei der Verwendung der Partizipialkonstruktion prägnanter. Bei Tabelle 5.52 sank die Stilqualität um 0,50 Punkte auf der Likert-Skala, während die Inhaltsqualität unverändert blieb. Neben der Bemängelung des Kongruenzfehlers beim Verb ‚is' kommentierte ein Bewerter „I would also take out ‚that are' for the sake of conciseness. ('Accessories specifically tailored to these speakers …')"

Erwartungsgemäß stiegen die Stil- und Inhaltsqualität in der Gruppe FR (MÜ falsch vor KS; richtig nach KS) und sanken in der Gruppe RF (MÜ richtig vor KS; falsch nach KS). In der Gruppe RF sanken die Stil- und Inhaltsqualität signifikant bei der Formulierung des Satzes ohne Partizipialkonstruktion (nach KS)

[44] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.90: Beispiel 57

Vor-KS	**Speziell auf diese Lautsprecher abgestimmtes Zubehör** erhalten Sie in unserem Webshop.
HMÜ Bing	Accessories tailored to these speakers specifically are available in our webshop.
Nach-KS	**Zubehör, das speziell auf diese Lautsprecher abgestimmt ist,** erhalten Sie in unserem Webshop.
HMÜ Bing	Accessories that is specifically tailored to these speakers are available in our webshop.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

aufgrund der aufgetretenen Fehler im Vergleich zu der fehlerfreien Übersetzung der Partizipialkonstruktion (vor KS), siehe Tabelle 5.91. In der Gruppe FR war der Anstieg der Stilqualität bei der Formulierung des Satzes ohne Partizipialkonstruktion (nach KS) insignifikant, während der Anstieg der Inhaltsqualität signifikant war, siehe Tabelle 5.91. Tabelle 5.92 veranschaulicht, dass die Stilqualität sank (− ,75 Punkte auf der Likert-Skala), obwohl die Inhaltsqualität durch den korrigierten Fehler stieg (+ 0,38 Punkte auf der Likert-Skala), Durch den behobenen lexikalischen Fehler in ‚fristgerecht' (LX.3: Wort wurde vor KS ausgelassen) stieg die Genauigkeit der MÜ nach KS. Zudem kritisierte ein Bewerter den Stil nach KS und beschrieb ihn als „wordy".

In der Gruppe RR (Übersetzung mit und ohne Passiversatz richtig) sanken die Stil- und Inhaltsqualität signifikant (z (N = 28) = − 3,540 / p < ,001) bzw. (z (N = 28) = 2,322 / p = ,020), siehe Tabelle 5.91. Eine richtige Übersetzung der Partizipialkonstruktion (vor KS) fanden die Bewerter prägnanter und idiomatisch. In Tabelle 5.93 kommentierte ein Bewerter die MÜ nach der Regalanwendung wie folgt: „'which go beyond this' too closely mimics the structure of the source text. I suggest ‚exceeding this' instead". Die Qualitätsdifferenz in diesem Beispiel betrug bei der Stilqualität − ,50 Punkte auf der Likert-Skala bzw. − ,38 Punkte auf der Likert-Skala bei der Inhaltsqualität.

Tabelle 5.91: „Partizipialkonst. verm." – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	37	,008	− 2,638
Differenz CQ (nach KS − vor KS)	37	,411	− ,823
Differenz allg. Q (nach KS − vor KS)	37	,507	− ,664
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	9	,107	− 1,612
Differenz CQ (nach KS − vor KS)	9	,021	− 2,314
Differenz allg. Q (nach KS − vor KS)	9	,050	− 1,958
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	23	< ,001	− 4,114
Differenz CQ (nach KS − vor KS)	23	,012	− 2,526
Differenz allg. Q (nach KS − vor KS)	23	< ,001	− 4,021
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	28	< ,001	− 3,540
Differenz CQ (nach KS − vor KS)	28	,020	− 2,322
Differenz allg. Q (nach KS − vor KS)	28	< ,001	− 3,618

5.4.6.8 Vergleich der AEM-Scores mit und ohne die Verwendung von Partizipialkonstruktionen sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores mit und ohne die Verwendung von Partizipialkonstruktionen zeigte sowohl mit TERbase als auch mit hLEPOR eine Verschlechterung der AEM-Scores (Abbildung 5.83).

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei ,164 (SD = ,214) und für die hLEPOR bei ,063 (SD = ,130) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben) (Abbildung 5.83). Die Differenzen (nach KS − vor KS) in TERbase und hLEPOR erwiesen sich als signifikant (z (N = 98) = − 6,238 / p < ,001) bzw. (z (N = 98) = − 4,366 / p < ,001).

Tabelle 5.92: Beispiel 58

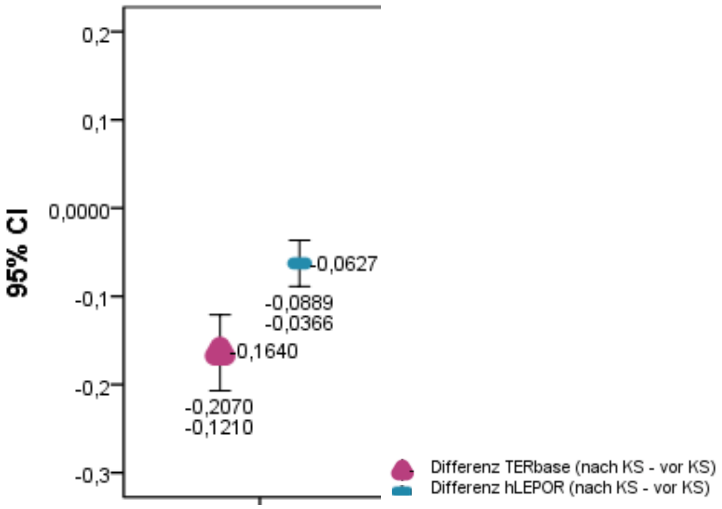
Vor-KS	Bei **fristgerecht erfolgten berechtigten Mängelrügen** ist der Lieferer zu einer kostenlosen Ersatzlieferung verpflichtet.
GNMÜ	In the case of XXX justified complaints, the supplier shall deliver replacement goods free of charge.
Nach-KS	Bei **berechtigten Mängelrügen**, **die fristgerecht erfolgen**, ist der Lieferer zu einer kostenlosen Ersatzlieferung verpflichtet.
GNMÜ	In the case of justified complaints which are made within the time limit, the supplier shall deliver replacement goods free of charge.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Tabelle 5.93: Beispiel 59

Vor-KS	**Alle darüberhinausgehenden Ansprüche** sind ausdrücklich von der Garantie ausgenommen.
HMÜ Bing	All claims that go beyond are expressly excluded from the guarantee.
Nach-KS	**Alle Ansprüche**, **die darüber hinausgehen**, sind ausdrücklich von der Garantie ausgenommen.
HMÜ Bing	All claims which go beyond this are expressly excluded from the guarantee.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.83: „Partizipialkonst. verm." – Mittelwert der Differenz der AEM-Scores

Dieses Ergebnis weist darauf hin, dass – nach der Auflösung der Partizipialkonstruktion (nach KS) – mehr Edits erforderlich waren.

5.4.6.8.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter mittlerer positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität (Tabelle 5.94).

Diesem Ergebnis zufolge standen die Qualitätsveränderungen der Humanevaluation und der automatischen Evaluation in relativem Einklang, denn der Qualitätsrückgang ging mit dem Rückgang der AEM-Scores einher.

5.4.6.9 Analyse der fünften Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der fünften analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (–)

Tabelle 5.94: „Partizipialkonst. verm." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	97	< ,001	,430	mittlerer Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	97	< ,001	,426	mittlerer Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[45]

Regel 5: Partizipialkonstruktionen vermeiden

Erster Analysefaktor: Vergleich der Fehleranzahl mit vs. ohne Partizipialkonstruktion

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

[45] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Resultat

Auf Regelebene: H0 wurde abgelehnt und somit H1 bestätigt. Die Fehleranzahl stieg signifikant, nachdem die Partizipialkonstruktion vermieden wurde.	**Anz.F. (+)**
Auf Regel- und MÜ-Systemebene: Bei Systran stieg die Fehleranzahl signifikant, nachdem die Partizipialkonstruktion vermieden wurde.	**Sy (+)**
Die Veränderungen der Fehleranzahl bei allen anderen Systemen waren nicht signifikant.	**Bi (+) Go (+) Lu (+) SD (–)**

Zweiter Analysefaktor

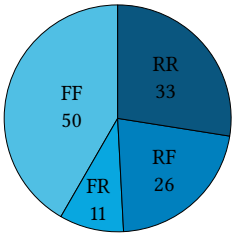


Abbildung 5.84: Aufteilung der Annotationsgruppen auf Regelebene

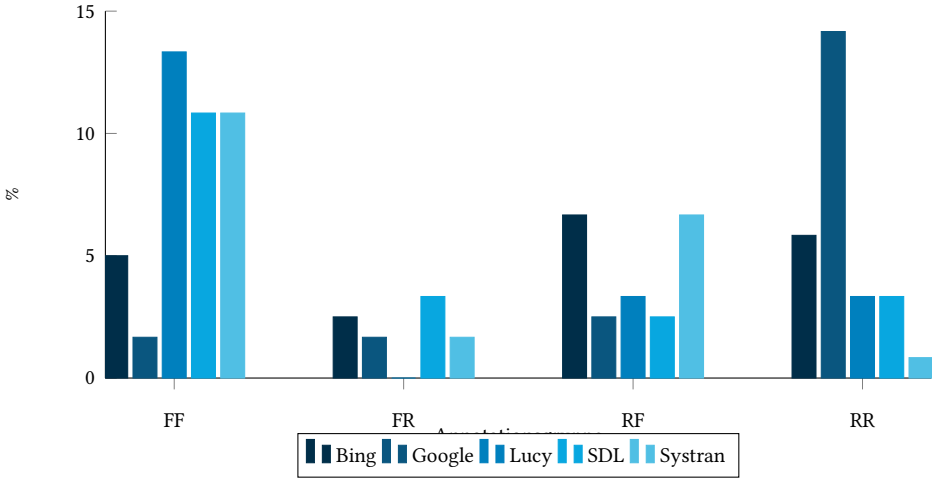


Abbildung 5.85: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Dritter Analysefaktor: Vergleich der Fehlertypen mit vs. ohne Partizipialkonstruktion

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Auf Regelebene:
H1 wurde für zwei Fehlertypen bestätigt.
Die Fehleranzahl von OR.1 „Zeichensetzung“ stieg signifikant, nachdem die Partizipialkonstruktion vermieden wurde.
Die Fehleranzahl von GR.10 „Wortstellungsfehler“ sank signifikant, nachdem die Partizipialkonstruktion vermieden wurde.

OR.1 (+)
GR.10 (–)

Auf Regel- und MÜ-Systemebene:
Bei Bing, SDL und Systran stieg die Fehleranzahl von OR.1 „Zeichensetzung“ signifikant, nachdem die Partizipialkonstruktion vermieden wurde.

OR.1 (+): Bi SD Sy

Bei SDL und Systran sank die Fehleranzahl von GR.10 „Wortstellungsfehler“ signifikant (nach KS).

GR.10 (–): SD Sy

Alle weiteren Veränderungen waren nicht signifikant.

Vierter Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne Partizipialkonstruktion

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
Für die Stilqualität wurde H0 abgelehnt und somit H1 bestätigt. **SQ (–)**

Für die Inhaltsqualität wurde H0 nicht abgelehnt und somit konnte H1 nicht bestätigt werden. **CQ (–)**

Auf Regel- und MÜ-Systemebene:
Die Stilqualität sank bei Bing, Google und Lucy signifikant (nach KS). **SQ (–): Bi Go Lu**

Die Inhaltsqualität stieg bei Systran signifikant (nach KS). **CQ (+): Sy**

Alle weiteren Qualitätsveränderungen waren nicht signifikant; generell sanken sowohl die Stil- als auch die Inhaltsqualität in den weiteren Fällen mit Ausnahme der Inhaltsqualität von SDL, die leicht anstieg.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde für drei Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und der Stilqualität sowie ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“, LX.4 „Zusätzliches Wort eingefügt“ und GR.10 „Wortstellungsfehler“ einzeln und der Differenz der Inhaltsqualität.

***neg* LX.4 <> SQ**

***neg* LX.3 <> CQ**
***neg* LX.4 <> CQ**
***neg* GR.10 <> CQ**

Auf Regel- und MÜ-Systemebene:

Bei Bing bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Wortstellungsfehler“ und der Differenz der Stilqualität sowie ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des SM.11 „Verwechslung des Sinns“ und der Differenz der Inhaltsqualität.

Bi
***neg* GR.10 <<>> SQ**
***neg* SM.11 <> CQ**

Bei SDL bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“, LX.4 „Zusätzliches Wort eingefügt“ und des GR.10 „Falsche Wortstellung“ einzeln und der Differenz der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des OR.1 „Falsche Zeichensetzung“, LX.3, LX.4 und GR.10 einzeln und der Differenz der Inhaltsqualität.

SD
***neg* LX.3 <<>> SQ**
***neg* LX.4 <<>> SQ**
***neg* GR.10 <<>> SQ**

***neg* OR.1 <<>> CQ**
***neg* LX.3 <<>> CQ**
***neg* LX.4 <<>> CQ**
***neg* GR.10 <<>> CQ**

Bei Systran bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“ und der Differenz der Inhaltsqualität.

Sy
***neg* LX.3 <<>> CQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne Partizipialkonstruktion auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde bei der Annotationsgruppe FF nur für die Stilqualität bestätigt: Die Stilqualität sank signifikant, nachdem die Partizipialkonstruktion vermieden wurde.	**SQ (–)**
Die Inhaltsqualität stieg nur minimal an.	**CQ (+)**
H1 wurde bei der Annotationsgruppe FR nur für die Inhaltsqualität bestätigt: Die Inhaltsqualität stieg nach der Vermeidung der Partizipialkonstruktion signifikant.	**CQ (+)**
Die Stilqualität stieg nicht signifikant an.	**SQ (+)**
Bei den Annotationsgruppen RF und RR sanken die Stil- und Inhaltsqualität signifikant (nach KS).	**SQ (–)** **CQ (–)**

Siebter Analysefaktor: Vergleich der AEM-Scores mit vs. ohne Partizipialkonstruktion

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt. Die AEM-Scores von TERbase und hLEPOR verschlechterten sich signifikant, nachdem die Partizipialkonstruktion vermieden wurde.	**TERbase (–)** **hLEPOR (–)**

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz des TERbase-Scores und der Differenz der allgemeinen Qualität sowie ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz des hLEPOR-Scores und der Differenz der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.7 SECHSTE REGEL: Passiv vermeiden

5.4.7.1 Überblick

Im Folgenden wird die KS-Regel „Passiv vermeiden" kurz beschrieben.[46] Zudem wird zusammenfassend und anhand von Beispielen demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Passiv vermeiden (tekom-Regel-Nr. S 501, S 502, S 503, S 504)

Nach diesen vier Regeln (tekom 2013: 79ff.) soll die Verwendung des Passivs vermieden und stattdessen sollen die Sätze im Aktiv formuliert werden.

Begründung: Die Passivkonstruktion ist oft nicht eindeutig und lässt den Handelnden unklar. Wenn der Täter genannt werden soll, eignet sich die

[46] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

Aktivformulierung. (ebd.: 79) Insbesondere bei Anweisungen, Sicherheits- und Warnhinweisen wird die Aktivkonstruktion empfohlen, damit dem Leser klar wird, wer die Handlung ausführt oder ausführen soll. Warnungen wirken motivierend, wenn sie im Aktiv formuliert sind, da der Leser direkt angesprochen wird. (ebd.: 81)

Umsetzungsmuster:
Vor KS: Satz formuliert im Passiv
Nach KS: Satz umformuliert im Aktiv

KS-Stelle
Vor KS: Form von ‚werden' + Partizip II
Nach KS: das Verb bzw. das Subjekt + das Verb, wenn das Subjekt im Passivsatz nicht enthalten war und erst im Aktivsatz hinzugefügt wurde

Die Subjekte können bei der MÜ semantisch falsch übersetzt werden, daher werden sie nur als Teil der KS-Stelle betrachtet, wenn sie im Passivsatz nicht existieren (vgl. folgende Beispiele).

Beispiele
Bei der Arbeit mit elektrischen Geräten sollte stets ein Sicherheitsstecker verwendet werden.

Bei der Arbeit mit elektrischen Geräten verwenden Sie stets einen Sicherheitsstecker.
Das Programm wird vom Hersteller wie folgt eingestellt.

Der Hersteller stellt das Programm wie folgt ein.

Aufteilung der Testsätze: In den deutschen Benutzerhandbüchern und Bedienungsanleitungen kommt häufig das Passiv in Kombination mit Modalverben vor, entsprechend besteht der Datensatz aus 20 Passivsätzen mit Modalverben und 4 Sätzen mit Vorgangspassiv ohne Modalverben.

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.7.2 Vergleich der Fehleranzahl beim Passiv vs. Aktiv

Die Fehleranzahl stieg um 30,5 % von 59 Fehlern im Falle der Verwendung von Passivkonstruktionen (M = ,49 / SD = ,733 / N = 120) auf 77 Fehler bei der Verwendung von Aktivkonstruktionen (M = ,64 / SD = ,914 / N = 120) (Abbildung

5.86). Der Mittelwert der Differenz (nach KS – vor KS) der Fehleranzahl pro Satz lag somit bei ,15 (SD = ,932) mit einem 95%-Konfidenzintervall zwischen einem Minimum von – ,02 (SD = ,731) und einem Maximum von ,32 (SD = 1,111) (Bootstrapping mit 1000 Stichproben) (Abbildung 5.87). Die Differenz (nach KS – vor KS) der Fehleranzahl war entsprechend nicht signifikant (z (N = 120) = – 1,688 / p = ,091).

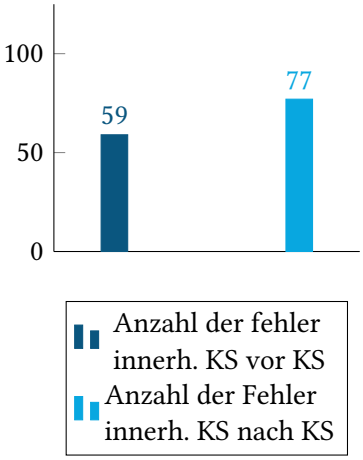


Abbildung 5.86: „Passiv verm." – Fehlersumme vor vs. nach KS

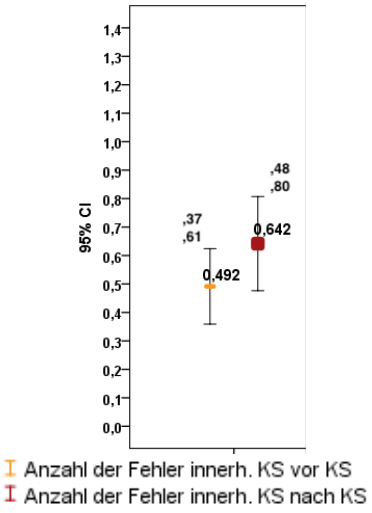


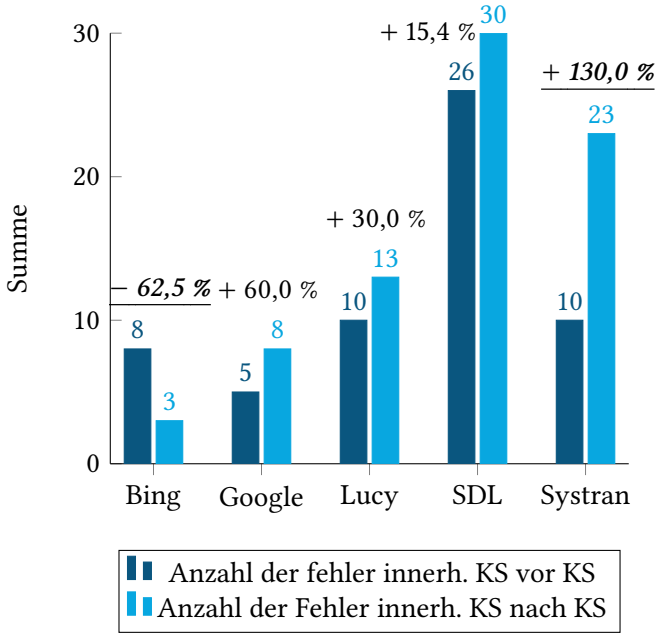
Abbildung 5.87: „Passiv verm." – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Die Fehlertypen, die erst nach der Regelanwendung auftraten, waren sehr unterschiedlich (mehr dazu unter §5.4.7.4). Daher lässt sich kein Muster ableiten, mit dessen Hilfe die Ursache des Anstiegs der Fehleranzahl interpretiert werden kann.

5.4.7.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Zwei signifikante gegensätzliche Reaktionen zeigten die beiden HMÜ-Systeme Bing und Systran: Während die Fehleranzahl bei Bing sank (Mdiff = – ,208; z (N = 24) = – 2,236 / p = ,025) stieg sie deutlich bei Systran (Mdiff = ,542; z (N = 24) = – 2,812 / p = ,005).

Unter den fünf analysierten Systemen stieg die Fehleranzahl nach der Verwendung der Aktivform bei allen Systemen mit Ausnahme von dem HMÜ-System Bing: Google (Mdiff = ,125); Lucy (Mdiff = ,125); SDL (Mdiff = ,167) (Abbildung



Signifikante Differenz vor vs. nach KS

Abbildung 5.88: „Passiv verm." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

5.88). Die Zunahme der Fehleranzahl bei Google, Lucy und SDL bei der Verwendung des Aktivs (nach KS) war nicht signifikant. Bei dem NMÜ-System Google Translate und dem SMÜ-System SDL ist eine mögliche Interpretation des (insignifikanten) Anstiegs der Fehleranzahl bei der Verwendung des Aktivs (nach KS), dass mehr Passiv- als Aktivsätze in den Trainingsdaten dieser Systeme enthalten sind. Die Veränderung in der Fehleranzahl im Falle des RBMÜ-Systems und der Hybridsysteme lässt sich im Rahmen einer Black-Box-Analyse nicht interpretieren.

5.4.7.3 Aufteilung der Annotationsgruppen

Die größte Annotationsgruppe bei dieser Regel war die Gruppe RR; knapp die Hälfte der Sätze wurde sowohl im Passiv als auch im Aktiv fehlerfrei übersetzt (Abbildung 5.89). An der zweiten Stelle kommt die Gruppe FF mit ca. 29 %; hier beinhalteten die MÜ in beiden Szenarien Fehler. Dann folgt die Gruppe RF mit ca. 13 %, in der die Sätze im Passiv, aber nicht im Aktiv, fehlerfrei übersetzt wur-

den. Zum Schluss kommt die Gruppe FR mit ca. 8 %, die nur im Aktiv fehlerfrei übersetzt werden konnte (Abbildung 5.89).

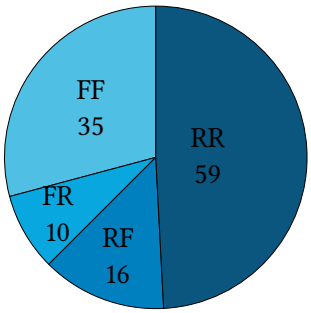


Abbildung 5.89: „Passiv verm." – Aufteilung der Annotationsgruppen

Im kommenden Abschnitt (§5.4.7.4) werden die Fehlertypen, die mit dem Passiv bzw. dem Aktiv verbunden sind, gegenübergestellt.

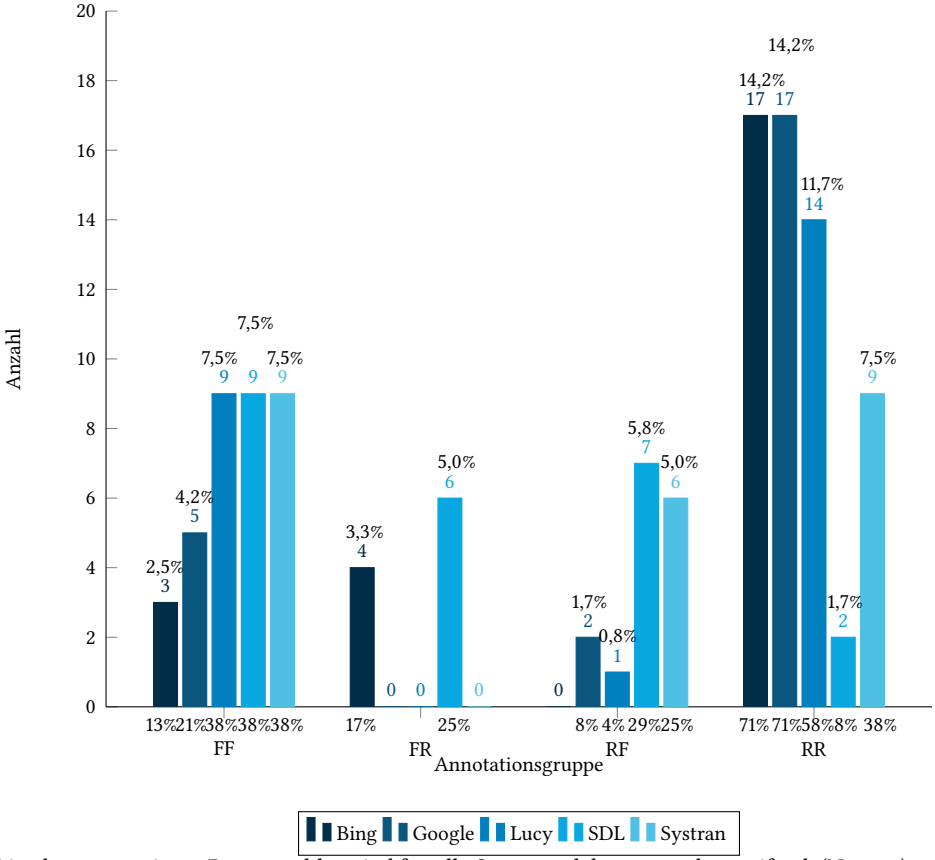
5.4.7.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Die größten Anteile bei drei MÜ-Systemen waren bei der Gruppe RR: 71 % der Sätze des NMÜ-Systems Google und des HMÜ-Systems Bing sowie 58 % des RBMÜ-Systems Lucy wurden sowohl im Passiv (vor KS) als auch im Aktiv fehlerfrei übersetzt (Abbildung 5.90). Gleichzeitig beinhalteten bei Lucy 38 % der Übersetzungen sowohl vor als auch nach der Anwendung der KS-Regel Fehler. Die Ergebnisse bei dem HMÜ-System Systran waren ebenfalls gemischt: 38 % bei der Gruppe RR und 38 % bei der Gruppe FF. Zudem waren 25 % der Übersetzungen von Systran im Passiv korrekt und nach der Anwendung des Aktivs falsch (Gruppe RF). Das SMÜ-System SDL erzielte auch ein heterogenes Ergebnis: 38 % in beiden Szenarien falsch (Gruppe FF); 29 % nur im Aktiv falsch (Gruppe RF) und 25 % nur im Passiv falsch (Gruppe FR).

Bei der Regel „Passiv vermeiden" zeigt ein genauer Einblick in den Datensatz, dass die Annotationsgruppe RF aus insgesamt 16 MÜ besteht, wobei die meisten Fälle dieser MÜ bei hauptsächlich zwei MÜ-Systemen vorkamen; und zwar 7 Fälle bei SDL (47 %) und 6 Fälle bei Systran (40 %) (Abbildung 5.90).

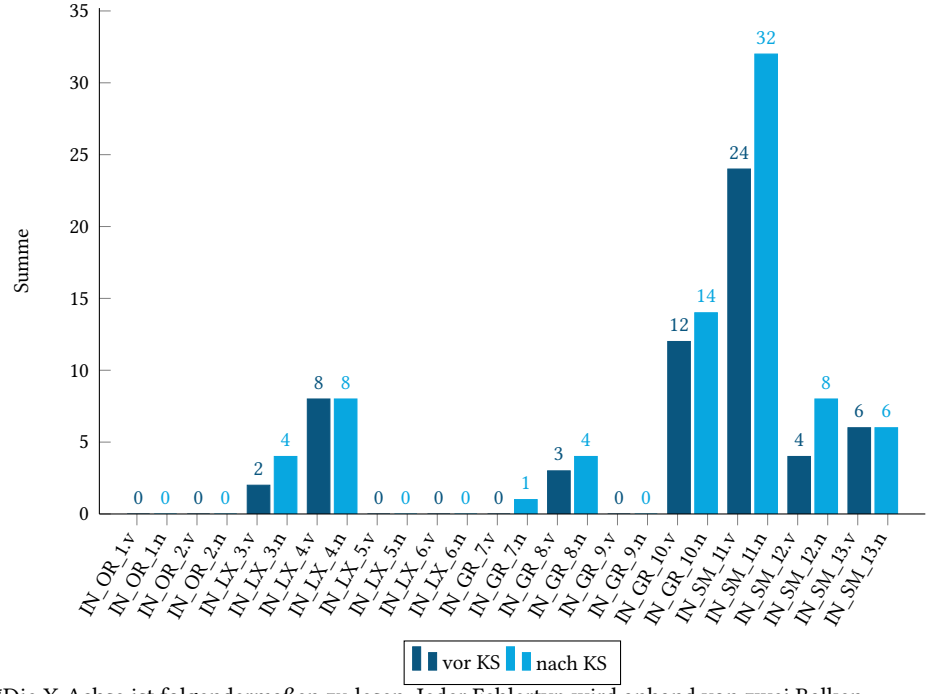
5.4.7.4 Vergleich der Fehlertypen beim Passiv vs. Aktiv

Die Fehleranzahl zeigte bei keinem bestimmten Fehlertyp eine deutliche Veränderung. Sowohl bei der Formulierung im Aktiv als auch im Passiv kommen unterschiedliche Fehler(typen) vor (Abbildung 5.91).



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.90: „Passiv verm.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler vor KS und der zweite die Summe der Fehler nach KS, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.91: „Passiv verm.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

Dementsprechend erwies sich der Unterschied in der Fehleranzahl bei keinem Fehlertyp als signifikant. Die am häufigsten aufgetretenen Fehler waren der Wortstellungsfehler (GR.10) und die Verwechslung des Sinns (SM.11), siehe Abbildung 5.91. Beide Fehlertypen stiegen im Aktiv (nach KS) leicht an. Nachfolgend zwei Beispiele der beiden Fehlertypen (Tabelle 5.95).

Tabelle 5.95: Beispiel 60

Vor-KS	Das Programm **wird** vom Hersteller wie folgt **eingestellt**.
HMÜ Systran	The program is stopped by the manufacturer as follows.
Nach-KS	Der Hersteller **stellt** das Programm wie folgt **ein**.
HMÜ Systran	The manufacturer stops the program as follows.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

In Tabelle 5.95 wird das Verb ‚einstellen' verwendet. Es handelt sich dabei um ein ambiges Verb, das u. a. als ‚stop' oder ‚set' übersetzt werden kann. Die Regel (Verwendung vom Passiv vs. Aktiv) kann bei solchen Fällen einen semantischen Fehler dieser Art nicht beeinflussen. Dies ist anders als der Wortstellungsfehler (GR.10) in Tabelle 5.96, bei dem man erwartet, dass Google Translate in der Lage sei, den Satz ohne Wortstellungsprobleme zu übersetzen.

Tabelle 5.96: Beispiel 61

Vor-KS	Flecken **sollten** so schnell wie möglich **behandelt werden**.
GNMÜ	Stains should be treated as soon as possible.
Nach-KS	Flecken **sollten Sie** so schnell wie möglich **behandeln**.
GNMÜ	Stains should treat you as soon as possible.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Im nächsten Abschnitt wird die Fehleranzahl der verschiedenen Systeme pro Fehlertyp gegenübergestellt.

5.4.7.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Betrachtung der Fehlertypen auf Systemebene zeigt (Abbildung 5.92), dass die Fehleranzahl bei den einzelnen Fehlertypen im Allgemeinen gering war bzw. sich meistens vor KS im Vergleich zu nach KS kaum veränderte.

Auf Systemebene waren die am meisten aufgetretenen Fehlertypen LX.4 „Zusätzliches Wort eingefügt“, GR.10 „Wortstellungsfehler“ (Tabelle 5.96) und SM.11 „Verwechslung des Sinns“ (Tabelle 5.95). Ein Beispiel für den Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ zeigt der folgende Fall (Tabelle 5.97):

Tabelle 5.97: Beispiel 62

Vor-KS	Bei der Arbeit mit elektrischen Geräten **sollte** stets ein Sicherheitsstecker **verwendet werden**.
HMÜ Systran	When working with electrical devices, a safety plug should always be used.
Nach-KS	Bei der Arbeit mit elektrischen Geräten **verwenden Sie** stets einen Sicherheitsstecker.
HMÜ Systran	When working with electrical devices, you always use a safety plug.

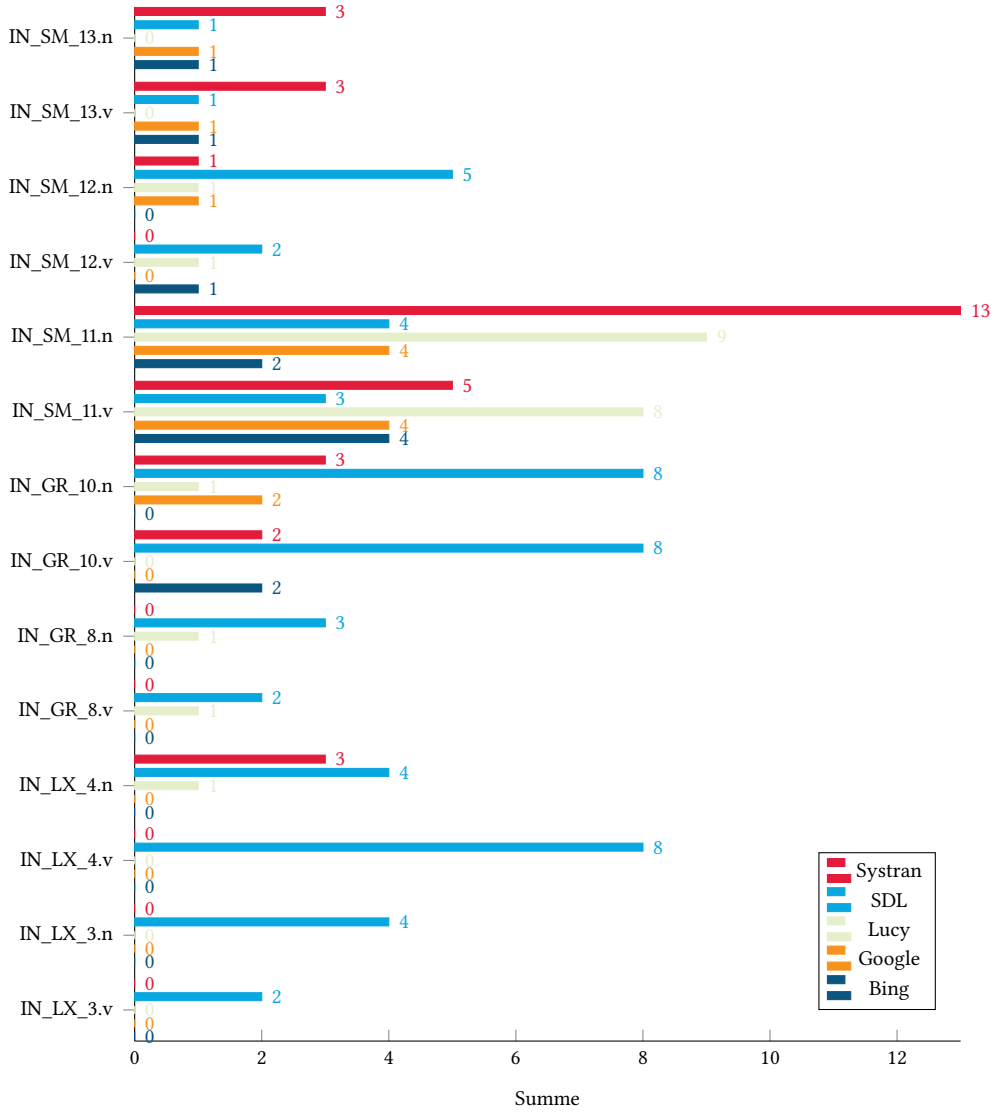
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

In Tabelle 5.97 handelt es sich um eine Sicherheitsanweisung. Bei dieser Art der technischen Kommunikation wird empfohlen, den Nutzer direkt anzusprechen, d. h. das Aktiv bzw. den Imperativ zu verwenden. Dennoch war Systran nicht in der Lage, die Imperativformulierung korrekt zu parsen.

Unabhängig von der Fehleranzahl und sogar bei fehlerfreien MÜ hat eine Formulierung im Passiv vs. Aktiv einen großen Einfluss auf den Stil. Daher werden im nächsten Abschnitt die Stil- und Inhaltsqualität analysiert und insbesondere bei der größten Annotationsgruppe RR näher betrachtet.

5.4.7.5 Vergleich der MÜ-Qualität beim Passiv vs. Aktiv sowie die Korrelation zwischen den Fehlertypen und der Qualität

Unter den neun analysierten Regeln kommt die Regel „Passiv vermeiden“ auf Platz eins mit dem höchsten Rückgang der Stil- und Inhaltsqualität nach der Re-



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 1, 2, 5, 6 und 9 bei gar keinem MÜ-System vor. Zudem kam der Fehlertyp 7 nur einmal jeweils bei einem MÜ-System vor.

Abbildung 5.92: „Passiv. verm.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

gelanwendung:[47] Die Stilqualität sank um 5,9 % (Mv = 4,41 / SDv = ,482 / Mn = 4,15 / SDn = ,506 / N = 83). Die Inhaltsqualität sank um 5,6 % (Mv = 4,62 / SDv = ,597 / Mn = 4,36 / SDn = ,923 / N = 83), siehe Abbildung 5.93. Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei − ,265 (SD = ,610) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,398 und einem Maximum von − ,132 und für die Inhaltsqualität bei − ,262 (SD = ,944) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,468 und einem Maximum von − ,056 (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.94. Die Differenzen (nach KS − vor KS) in der Stil- und Inhaltsqualität erwiesen sich als hochsignifikant (z (N = 83) = − 6,235 / p < ,001) bzw. (z (N = 83) = − 4,740 / p < ,001).

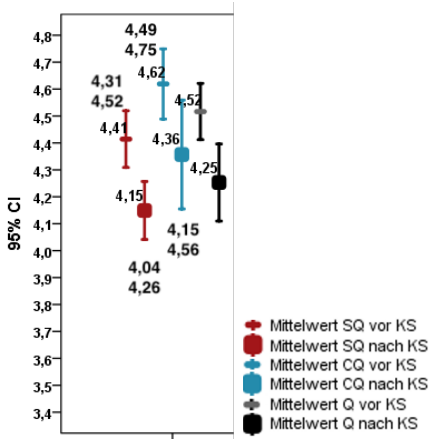


Abbildung 5.93: „Passiv verm." – Mittelwerte der Qualität vor und nach KS

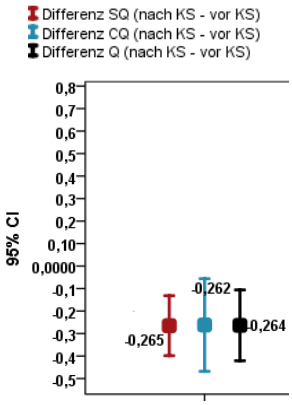


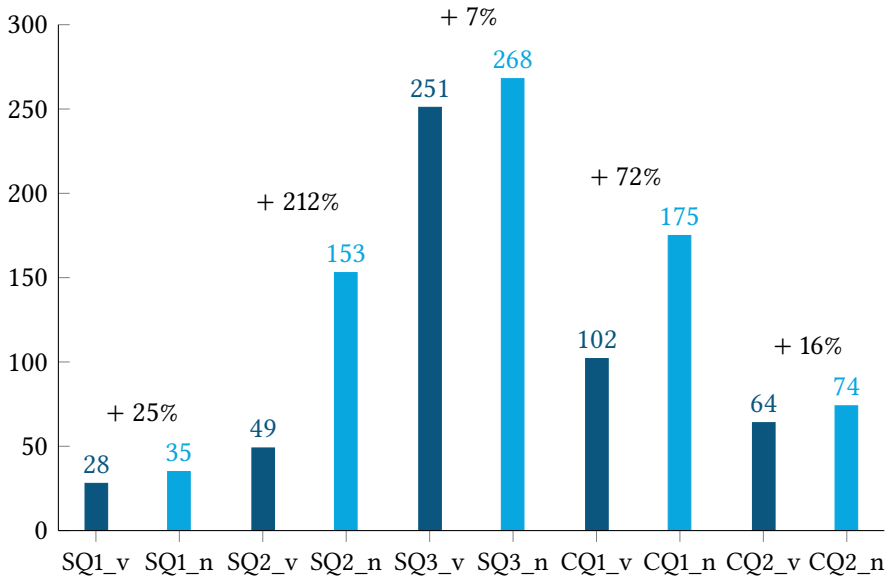
Abbildung 5.94: „Passiv verm." – Mittelwert der Qualitätsdifferenzen

Die Humanevaluation deckt den Grund des Qualitätsrückgangs auf. Wie Abbildung 5.95 zeigt, verschlechtern sich alle Qualitätskriterien nach der Verwendung einer Aktivformulierung (nach KS). Hierbei spielten das zweite Stilqualitätskriterium (SQ2 – Stilistische Adäquatheit)[48] zusammen mit dem ersten Inhaltsqualitätskriterium (CQ1 – Genauigkeit) die wesentliche Rolle bei der Qualitätsveränderung.

In Tabelle 5.98 sanken die Stilqualität (− ,63 Punkte auf der Likert-Skala) und die Inhaltsqualität (− ,25 Punkte auf der Likert-Skala).

[47] Definitionen der Qualität unter §4.5.5.1.

[48] Stilistische Adäquatheit im Sinne von Hutchins & Somers (1992: 163) ist „the extent to which the translation uses the language appropriate to its content and intention". Mehr zu den Definitionen der Qualität unter §4.5.5.1.



SQ1: Ü ist **nicht** korrekt bzw. **nicht** klar dargestellt, d. h. nicht orthografisch
SQ2: Ü ist **nicht** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer **nicht** zum Handeln, zieht **nicht** seine Aufmerksamkeit an usw.
SQ3: Ü klingt **nicht** natürlich bzw. **nicht** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext **nicht** exakt wieder.
CQ2: Ü ist **nicht** leicht zu verstehen, d. h. **nicht** gut formuliert bzw. dargestellt.

Abbildung 5.95: „Passiv verm.“ – Vergleich der Qualitätskriterien

Tabelle 5.98: Beispiel 63

Vor-KS	Die akustischen Signale **können** je nach Gerät **umprogrammiert werden**.
GNMÜ	Depending on the device, the acoustic signals can be reprogrammed.
Nach-KS	Die akustischen Signale **können Sie** je nach Gerät **umprogrammieren**.
GNMÜ	Depending on the device, you can reprogram the acoustic signals.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Nach der Regelanwendung wird der Nutzer durch das Aktiv direkt angesprochen. Dies fanden die Bewerter stilistisch nicht erforderlich (SQ2). Einer der Bewerterkommentare lautete: „It's not common to directly address the reader as ‚you' in these contexts. I'd recommend the passive voice instead." Der geringe Punktabzug bei der Inhaltsqualität erfolgte aufgrund der Unsicherheit bei der Genauigkeit der Übersetzung (CQ1) des Verbs ‚umprogrammieren' als ‚reprogram'. Hierbei wünschten sich die Bewerter für eine adäquate Übersetzung mehr Kontextinformationen.

5.4.7.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung bei der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS – Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS – Q. vor KS) zusammenhängt. Bei der Stilqualität gab es zwei signifikante mittlere negative Korrelationen zwischen der Differenz in den Fehlertypen LX.4 „Zusätzliches Wort eingefügt" und GR.10 „Wortstellungsfehler" einzeln und der Differenz in der Stilqualität. Die weiteren signifikanten Korrelationen zwischen der Differenz in den Fehlertypen LX.3 „Wort ausgelassen" und SM.11 „Verwechslung des Sinns" einzeln und der Differenz in der Stilqualität waren schwache negative Korrelationen. (siehe Tabelle 5.99)

Bei der Inhaltsqualität gab es drei signifikante mittlere negative Korrelationen zwischen der Differenz in den Fehlertypen GR.10, SM.11 und SM.12 „Falsche Wahl" einzeln und der Differenz in der Inhaltsqualität. Die weiteren signifikanten Korrelationen zwischen der Differenz in den Fehlertypen LX.3 und LX.4 einzeln und der Differenz in der Inhaltsqualität waren schwache negative Korrelationen. (siehe Tabelle 5.99)

Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht erwiesen werden.

Bei dieser Regel stiegen die vorgekommenen Fehlertypen entweder leicht oder blieben nach der Regelanwendung unverändert. Insgesamt sanken sowohl die Stilqualität als auch die Inhaltsqualität signifikant. In Tabelle 5.100 beobachten wir, wie sich der Wortstellungsfehler in ‚you can connect' nach der Regelanwendung auf die MÜ auswirkt.

Nach der Formulierung des Satzes im Aktiv, trat der Wortstellungsfehler auf. Daraufhin sanken die Stilqualität um 1,25 Punkte und die Inhaltsqualität um 1,63 Punkte auf der Likert-Skala.

Tabelle 5.99: „Passiv verm.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	83	,014	− ,269
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	83	,001	− ,371
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	83	< ,001	− ,431
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	83	,038	− ,228
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	83	,271	− ,122
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	83	,007	− ,293
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	83	,008	− ,290
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	83	< ,001	− ,473
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	83	,001	− ,354
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	83	,004	− ,315
Differenz allg. Q (nach KS − vor KS)			
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	83	,008	− ,290
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	83	,001	− ,349
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	83	< ,001	− ,441
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	83	,004	− ,312
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	83	,005	− ,302

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Tabelle 5.100: Beispiel 64

Vor-KS	Durch diese Öffnung **kann** der Stecker mit dem Regler **verbunden werden**.
SMÜ SDL	Through this opening, the plug can be connected to the controller.
Nach-KS	Durch diese Öffnung **können Sie** den Stecker mit dem Regler **verbinden**.
SMÜ SDL	Through this opening, the plug you can connect to the controller.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.7.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.96 zeigt, sanken grundsätzlich die Stil- und Inhaltsqualität nach der Regelanwendung bei allen Systemen mit der Ausnahme eines leichten Anstiegs der Inhaltsqualität bei dem HMÜ-System Bing:

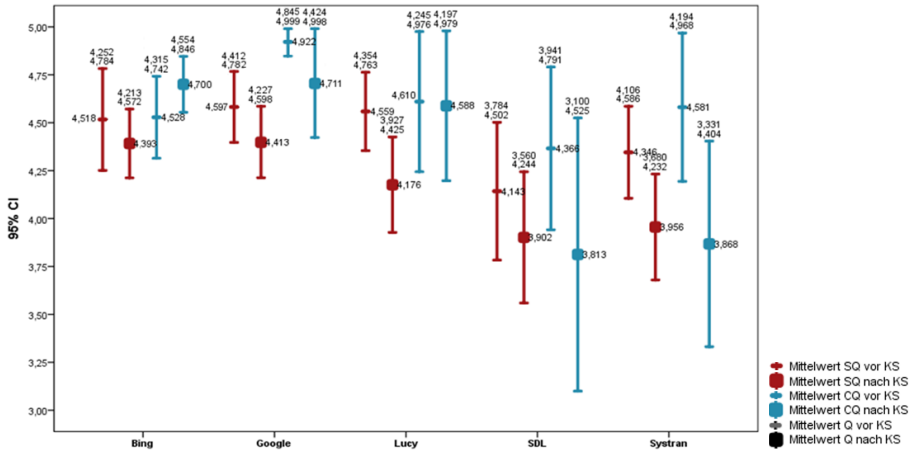


Abbildung 5.96: „Passiv verm." – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Die signifikanten Rückgänge der Stilqualität wurden bei zwei Systemen registriert: dem RBMÜ-System Lucy (– 8,4 %) und dem HMÜ-System Systran (– 9,0 %). Einen signifikanten Rückgang der Inhaltsqualität gab es nur bei Systran (– 15,6 %). (siehe Tabelle 5.101)

Tabelle 5.101: „Passiv verm.“ – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	16	,130	− 1,515	16	,138	− 1,485	16	,909	− ,114
Google	19	,138	− 1,483	19	,200	− 1,280	19	,114	− 1,582
Lucy	17	,007	− 2,700	17	1,000	,000	17	,023	− 2,282
SDL	14	,329	− ,975	14	,184	− 1,329	14	,249	− 1,154
Systran	17	,031	− 2,155	17	,009	− 2,603	17	,008	− 2,643

p: Signifikanz z: Teststatistik nicht signifikant ($p \geq 0{,}05$)

Der signifikante Rückgang der Stil- und Inhaltsqualität bei dem HMÜ-System Systran kam zum großen Teil dadurch, dass das Subjekt ‚Sie‘ in der aktiven Version (nach KS) als ‚they‘ anstatt ‚you‘ übersetzt wurde (Semantikfehler SM.11 „Verwechslung des Sinns“). Wie Tabelle 5.102 zeigt, war dieser Semantikfehler der einzige Fehler nach der Anwendung der Regel.

Tabelle 5.102: Beispiel 65

Vor-KS	**Der EIN/AUS-Schalter** kann **komfortabel mit dem Fuß** betätigt werden.
HMÜ Systran	The ON/OFF switch can be operated conveniently with your foot.
Nach-KS	Sie können **den EIN/AUS-Schalter komfortabel mit dem Fuß** betätigen.
HMÜ Systran	They can operate the ON/OFF switch conveniently with your foot.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Bei diesem Satz sank nach der Regelanwendung die Stilqualität (− ,25 auf der Likert-Skala) und die Inhaltsqualität (− 1,63 auf der Likert-Skala).

5.4.7.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwies sich nur bei dem SMÜ-System SDL ein signifikanter starker negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.3 „Wort ausgelassen“, LX.4 „Zusätzliches Wort eingefügt“ und GR.10 „Wortstellungsfehler“ einzeln und der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.3 und GR.10 einzeln und der Inhaltsqualität (siehe Tabelle 5.103).

Tabelle 5.103: „Passiv verm.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	SDL		
	N	**p**	**ρ**
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl **LX.3 „W. fehlt“**	14	,048	− ,536
Diff. der Anzahl **LX.4 „W. extra“**	14	,028	− ,584
Diff. der Anzahl **GR.10 „Wortst.“**	14	< ,001	− ,807
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl **LX.3 „W. fehlt“**	14	,032	− ,573
Diff. der Anzahl **LX.4 „W. extra“**	14	,052	− ,528
Diff. der Anzahl **GR.10 „Wortst.“**	14	,014	− ,638
Differenz Q (nach KS − vor KS)			
Diff. der Anzahl **LX.3 „W. fehlt“**	14	,033	− ,572
Diff. der Anzahl **LX.4 „W. extra“**	14	,036	− ,564
Diff. der Anzahl **GR.10 „Wortst.“**	14	,006	− ,697

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (ρ ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

In Tabelle 5.104 erschienen der lexikalische Fehlertyp LX.3 „Wort ausgelassen“ (das Subjekt ‚you‘ sowie das Verb ‚treat‘ wurden ausgelassen) sowie der Wortstellungsfehler GR.10 (in ‚Stains should‘) nach der Verwendung des Aktivs (Nach KS).

Tabelle 5.104: Beispiel 66

Vor-KS	Flecken **sollten** so schnell wie möglich **behandelt werden**.
SMÜ SDL	Stains should be treated as soon as possible.
Nach-KS	Flecken **sollten Sie** so schnell wie möglich **behandeln**.
SMÜ SDL	XXX Stains should XXX as soon as possible.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Daraufhin sanken die Stilqualität (um 1,75 Punkte) und die Inhaltsqualität (um 3,50 Punkte auf der Likert-Skala) nach der Regelanwendung deutlich.

5.4.7.6 Vergleich der MÜ-Qualität beim Passiv vs. Aktiv auf Annotationsgruppenebene

Die MÜ-Qualität[49] sank nach der Regalanwendung in allen Annotationsgruppen mit zwei Ausnahmen: In der Gruppe FR (Passivsatz wurde falsch übersetzt; Aktivsatz wurde richtig übersetzt) stiegen die Stil- und Inhaltsqualität leicht; in der Gruppe RR (Satz wurde im Passiv und im Aktiv richtig übersetzt) stieg nur die Inhaltsqualität minimal (Abbildung 5.97).

In der Gruppe FF (Satz wurde im Passiv und im Aktiv falsch übersetzt) sank die Stilqualität signifikant (z (N = 25) = − 2,566 / p = ,010) und die Inhaltsqualität nicht signifikant (z (N = 25) = − 1,351 / p = ,177), siehe Tabelle 5.105.

Tabelle 5.106 zeigt, dass obwohl der semantische Fehler (in der Übersetzung des Verbs ‚einstellen‘ als ‚stop‘ anstatt ‚set‘) sich in beiden Szenarien (vor und nach KS) wiederholt, die Bewerter die Passivformulierung nicht kritisierten. Im Gegenteil – eine der empfohlenen Korrekturen war eine Passivformulierung „The program *has been set up* by the manufacturer as follows“. Entsprechend sank bei der Verwendung des Aktivs die Stilqualität um 0,38 Punkte und die Inhaltsqualität um 0,13 Punkte auf der Likert-Skala. Eine solche Empfehlung unterliegt der Einschätzung der Bewerter, inwiefern die Handlung mithilfe des Passivs in den Vordergrund gerückt werden sollte.

[49] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.105: „Passiv verm.“ – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS – vor KS)	25	,010	– 2.566
Differenz CQ (nach KS – vor KS)	25	,177	– 1,351
Differenz allg. Q (nach KS – vor KS)	25	,019	– 2,345
Annotationsgruppe FR			
Differenz SQ (nach KS – vor KS)	6	,114	– 1,581
Differenz CQ (nach KS – vor KS)	6	,080	– 1,753
Differenz allg. Q (nach KS – vor KS)	6	,075	– 1,782
Annotationsgruppe RF			
Differenz SQ (nach KS – vor KS)	12	,003	– 2,940
Differenz CQ (nach KS – vor KS)	12	,002	– 3,062
Differenz allg. Q (nach KS – vor KS)	12	,002	– 3,061
Annotationsgruppe RR			
Differenz SQ (nach KS – vor KS)	40	,061	– 1,876
Differenz CQ (nach KS – vor KS)	40	,427	– ,794
Differenz allg. Q (nach KS – vor KS)	40	,383	– ,872

Tabelle 5.106: Beipiel 67

Vor-KS	Das Programm **wird** vom Hersteller wie folgt **eingestellt**.
HMÜ Systran	The program is stopped by the manufacturer as follows.
Nach-KS	Der Hersteller **stellt** das Programm wie folgt **ein**.
HMÜ Systran	The manufacturer stops the program as follows.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

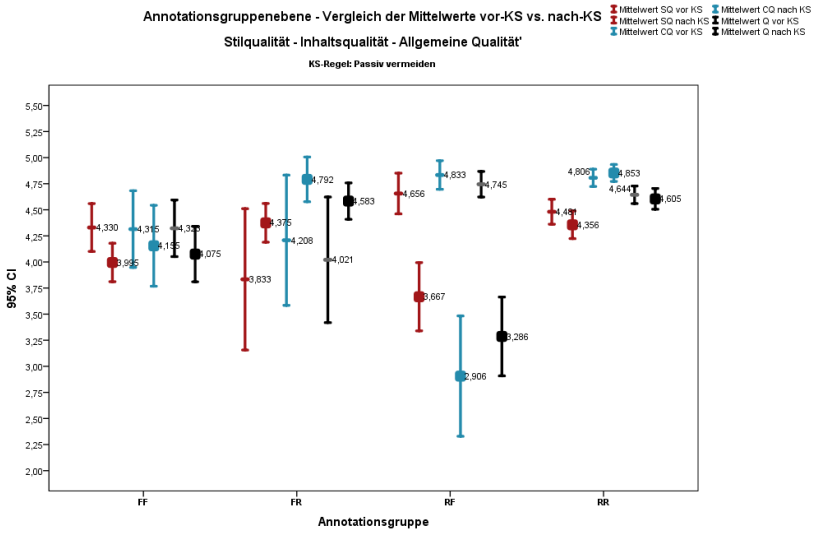


Abbildung 5.97: „Passiv verm." – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

Erwartungsgemäß stiegen die Stil- und Inhaltsqualität in der Gruppe FR (MÜ falsch vor KS; richtig nach KS), dennoch war der Anstieg nicht signifikant, siehe Abbildung 5.97 und Tabelle 5.105, denn diese Gruppe war relativ klein (8 %; 10 Sätze, siehe Abbildung 5.89).

In der Gruppe RF (MÜ richtig vor KS; falsch nach KS) sanken die Stil- und Inhaltsqualität signifikant, siehe Tabelle 5.105. Dieses Ergebnis ist nachvollziehbar, denn die Humanevaluation zeigte, dass eine richtig übersetzte Passivformulierung stilistisch und inhaltlich besser als eine falsche Aktivformulierung ist.

In der Gruppe RR (Satz wurde im Passiv und im Aktiv richtig übersetzt) sank die Stilqualität leicht und die Inhaltsqualität stieg minimal. Eine insignifikante Veränderung der Qualität bei dieser Gruppe zeigt, dass eine Aktivformulierung nicht unbedingt vorteilhaft ist. In Tabelle 5.107 empfahlen die Bewerter zur Verbesserung des Stils, den Satz im Passiv zu formulieren; einer der Bewerterkommentare lautete: „remove 'you' and rewrite as passive: 'The configuration of the module can be exported…'". Sie hielten somit eine direkte Anrede des Lesers nicht für erforderlich; dies hängt potenziell mit der Möglichkeitsformulierung (in ‚can be exproted') zusammen, die im Satz ausgedrückt werden soll.

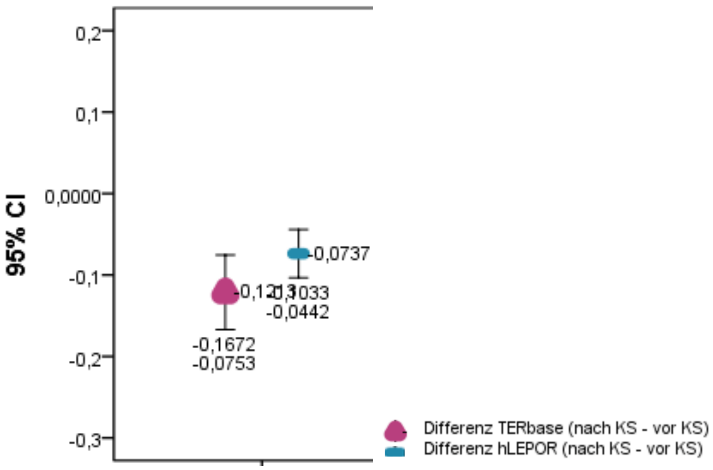
Tabelle 5.107: Beispiel 68

Vor-KS	Die Konfigurierung des Moduls **kann** in eine Datei **exportiert werden**.
HMÜ Bing	The configuration of the module can be exported to a file.
Nach-KS	**Sie können** die Konfigurierung des Moduls in eine Datei **exportieren**.
HMÜ Bing	You can export the configuration of the module to a file.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.7.7 Vergleich der AEM-Scores beim Passiv vs. Aktiv sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores nach der Verwendung des Aktivs (nach KS) zeigte sowohl mit TERbase als auch mit hLEPOR eine Verschlechterung der AEM-Scores.



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.98: „Passiv verm." – Mittelwert der Differenz der AEM-Scores

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei ,12 (SD = ,210) und für die hLEPOR bei ,074 (SD = ,135) mit einem

95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.78. Die Differenzen (nach KS − vor KS) in TERbase und hLEPOR erwiesen sich als hochsignifikant (z (N = 83) = − 4,717 / p < ,001) bzw. (z (N = 83) = − 4,400 / p < ,001). Dieses Ergebnis weist darauf hin, dass für die Korrektur der Aktivsätze mehr Edits erforderlich wären.

5.4.7.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Wie der vierte Analysefaktor (§5.4.7.5) zeigte, sank die Qualität nach der Regelanwendung signifikant. Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter starker positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität. Nach der Verwendung der Aktivformulierung (nach KS) verschlechterten sich die Scores der beiden AEMs und es war ein Qualitätsabfall zu bemerken.

Tabelle 5.108: „Passiv verm." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	83	< ,001	,504	starker Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	83	< ,001	,553	starker Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Diese signifikante positive Korrelation deutet darauf hin, dass die in der Humanevaluation und automatischen Evaluation festgestellten Qualitätsveränderungen übereinstimmen bzw. sich gegenseitig bestätigen.

5.4.7.8 Analyse der sechsten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der sechsten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (−) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[50]

Regel 6: Passiv vermeiden

Erster Analysefaktor: Vergleich der Fehleranzahl beim Passiv vs. Aktiv

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene: H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. Die Fehleranzahl stieg nach der Verwendung des Aktivs (nach KS), allerdings war der Anstieg nicht signifikant.	**Anz.F. (+)**
Auf Regel- und MÜ-Systemebene: Bei Bing sank die Fehleranzahl signifikant, nach der Verwendung des Aktivs (nach KS). Hingegen stieg bei Systran die Fehleranzahl.	**Bi (−)** **Sy (+)**
Bei den anderen drei Systemen stieg die Fehleranzahl nicht signifikant (nach KS).	**Go (+)** **Lu (+)** **SD (+)**

[50] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Zweiter Analysefaktor

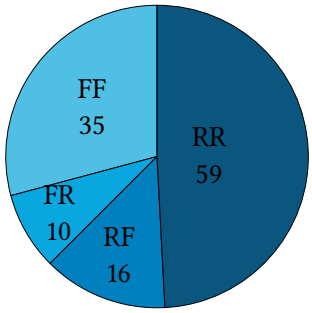


Abbildung 5.99: Aufteilung der Annotationsgruppen auf Regelebene

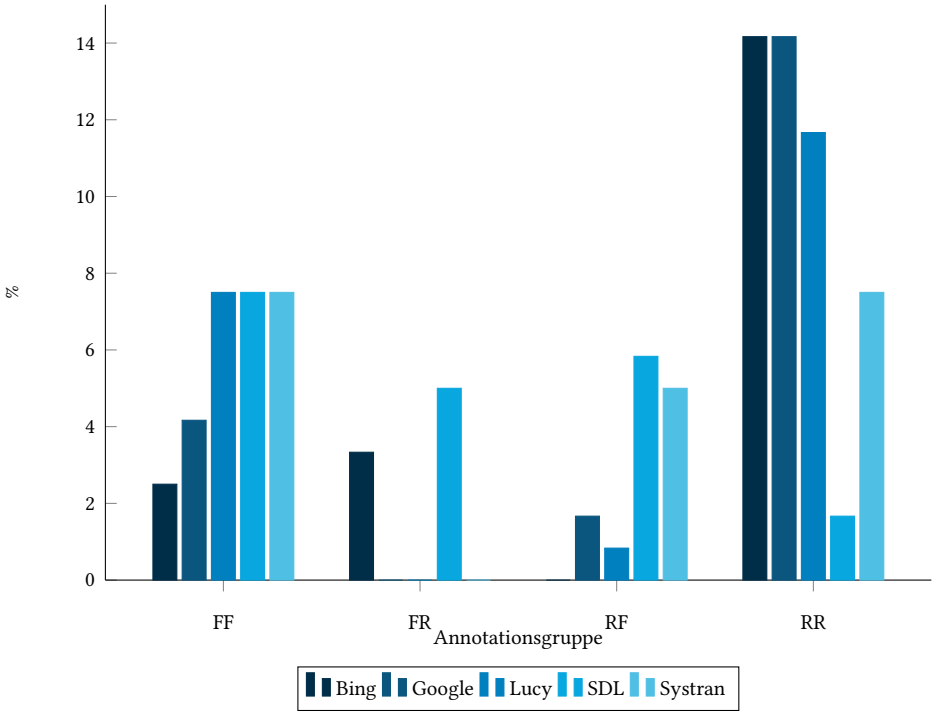


Abbildung 5.100: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Dritter Analysefaktor: Vergleich der Fehlertypen beim Passiv vs. Aktiv

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden.
Der Unterschied in der Fehleranzahl erwies sich bei keinem der Fehlertypen als signifikant.
Auf Regel- und MÜ-Systemebene:
Es gab bei keinem der Systeme signifikante Veränderungen in den Fehlertypen. Die Fehleranzahl bei den einzelnen Fehlertypen fiel im Allgemeinen gering aus bzw. veränderte sich meistens vor vs. nach KS kaum.

Vierter Analysefaktor: Vergleich der MÜ-Qualität beim Passiv vs. Aktiv

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt. **SQ (–)**
Sowohl die Stil- als auch die Inhaltsqualität sanken signifikant. **CQ (–)**
Auf Regel- und MÜ-Systemebene:
Die Stilqualität sank bei Lucy und Systran signifikant (nach KS). **SQ (–): Lu Sy**

Die Inhaltsqualität sank bei Systran signifikant (nach KS). **CQ (–): Sy**

Alle weiteren Qualitätsveränderungen waren nicht signifikant; generell sanken sowohl die Stil- als auch die Inhaltsqualität in den weiteren Fällen mit Ausnahme der Inhaltsqualität von Bing, die leicht anstieg.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS – vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde für vier Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und GR.10 „Wortstellungsfehler“ einzeln und der Stilqualität sowie ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „GR – Wortstellungsfehler“, SM.11 „Verwechslung des Sinns“ und SM.12 „Falsche Wahl“ einzeln und der Differenz der Inhaltsqualität.

***neg* LX.4 <> SQ**
***neg* GR.10 <> SQ**

***neg* GR.10 <> CQ**
***neg* SM.11 <> CQ**
***neg* SM.12 <> CQ**

Auf Regel- und MÜ-Systemebene:

Bei SDL bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“, LX.4 „Zusätzliches Wort eingefügt“ und GR.10 „Wortstellungsfehler“ einzeln und der Differenz der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.3 „Wort ausgelassen“ und GR.10 „Wortstellungsfehler“ einzeln und der Differenz der Inhaltsqualität.

SD
***neg* LX.3 <<>> SQ**
***neg* LX.4 <<>> SQ**
***neg* GR.10 <<>> SQ**

***neg* LX.3 <<>> CQ**
***neg* GR.10 <<>> CQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität beim Passiv vs. Aktiv auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde bei der Annotationsgruppe FF nur für die Stilqualität bestätigt:

Die Stilqualität sank signifikant bei der Aktivformulierung (nach KS).	**SQ (–)**
Der Rückgang der Inhaltsqualität war gering.	**CQ (–)**
Bei der Annotationsgruppen RF sanken die Stil- und Inhaltsqualität signifikant (nach KS).	**SQ (–)** **CQ (–)**
Bei der Annotationsgruppe FR stiegen die Qualitätswerte leicht.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe RR sank die Stilqualität leicht und die Inhaltsqualität stieg minimal.	**SQ (–)** **CQ (+)**

Siebter Analysefaktor: Vergleich der AEM-Scores beim Passiv vs. Aktiv

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Die AEM-Scores von TERbase und hLEPOR verschlechterten sich signifikant im Falle der Aktivformulierung (nach KS).

TERbase (–)
hLEPOR (–)

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz des TERbase-Scores und der Differenz der allgemeinen Qualität sowie ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz des hLEPOR-Scores und der Differenz der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.8 SIEBTE REGEL: Konstruktionen mit „sein + zu + Infinitiv“ vermeiden

5.4.8.1 Überblick

Im Folgenden wird die KS-Regel „Konstruktionen mit ‚sein + zu + Infinitiv‘ vermeiden“ kurz beschrieben.[51] Zudem wird zusammenfassend und anhand von Beispielen demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Konstruktionen mit „sein + zu + Infinitiv“ vermeiden (tekom-Regel-Nr. S 511)

Nach dieser Regel (tekom 2013: 86) soll die Passiversatzkonstruktion „sein + zu + Infinitiv“ bei Anweisungen vermieden werden. Stattdessen sollen diese durch einen Infinitiv oder die direkte Anrede formuliert werden.

Begründung: Diese Passiversatzkonstruktion ist umständlich und der Leser wird nicht direkt angesprochen. Der Infinitiv bzw. eine direkte Anrede fördert die schnelle und richtige Umsetzung der Handlung (ebd.).

Umsetzungsmuster:

Vor KS: Der Satz ist mit einem Passiversatz (sein + zu + Infinitiv) formuliert.
Nach KS: zwei Varianten sind möglich

– Imperativ am Satzende

– Imperativ am Satzanfang

Beide Varianten wurden bei der Übersetzung mit MÜ-Systemen getestet. Da die erste Variante mit mehr MÜ-Fehlern verbunden war, wurde entschieden, die zweite Variante als Umsetzungsmuster zu verwenden.

Wenn der Satz vor KS mit ‚so‘ formuliert ist, wurde ‚so‘ aus stilistischen Gründen nach KS entfernt.

KS-Stelle

Vor KS: sein + zu + Infinitiv
Nach KS: Imperativ + Subjekt

Beispiele

Die Herstellerangaben sind stets zu beachten .

[51] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

Beachten Sie stets die Herstellerangaben.

Ist ein mehrstufiges Modul parametriert, so sind die externen Kontakte zu verriegeln.

Ist ein mehrstufiges Modul parametriert, verriegeln Sie die externen Kontakte.

Aufteilung der Testsätze: Der Datensatz besteht aus 24 verschieden Verben im Passiversatz, die an unterschiedlichen Stellen in den Sätzen platziert sind.

Im Weiteren werden die Ergebnisse der einzelnen Analysefaktoren demonstriert.

5.4.8.2 Vergleich der Fehleranzahl mit vs. ohne die Konstruktion „sein + zu + Infinitiv“

Die Fehleranzahl sank um mehr als ein Drittel, nämlich um 37,2 %, von 86 Fehlern bei der Verwendung des Passiversatzes mit „sein + zu + Infinitiv“ (M = ,72 / SD = ,881 / N = 120) auf 54 Fehler bei der Vermeidung des Passiversatzes (M = ,45 / SD = ,765 / N = 120), siehe Abbildung 5.101. Knapp 80 % (19 von 24) der analysierten Sätze wurden von mindestens einem MÜ-System falsch übersetzt und mithilfe der KS-Regel korrigiert. Der Mittelwert der Differenz (nach KS − vor KS) der Fehleranzahl pro Satz lag somit bei − ,27 (SD = 1,098) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,47 (SD = ,864) und einem Maximum von − ,07 (SD = 1,311) (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.102. Die Differenz (nach KS − vor KS) der Fehleranzahl erwies sich als signifikant (z (N = 120) = − 3,059 / p = ,002).

Die Sätze, die falsch übersetzt wurden, haben keine gemeinsamen Eigenschaften. Sie sind unterschiedlich lang und der Passiversatz war unterschiedlich platziert. Es scheint, dass die MÜ-Systeme durch das zusätzliche Verb (sind) im Passiversatz irregeleitet werden und dadurch Schwierigkeiten haben, den Satz korrekt zu parsen. In Tabelle 5.109 konnte das Verb nicht richtig übersetzt werden (‚is … turn off‘). Durch die Regelanwendung konnte dieses Problem gelöst werden.

5.4.8.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Ein genauer Einblick bei den einzelnen MÜ-Systemen zeigt, dass die Fehleranzahl bei zwei MÜ-Systemen sank (Abbildung 5.103).

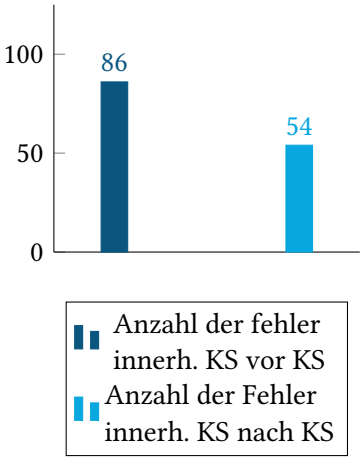


Abbildung 5.101: „Passiversatz verm.“ – Fehlersumme vor vs. nach KS

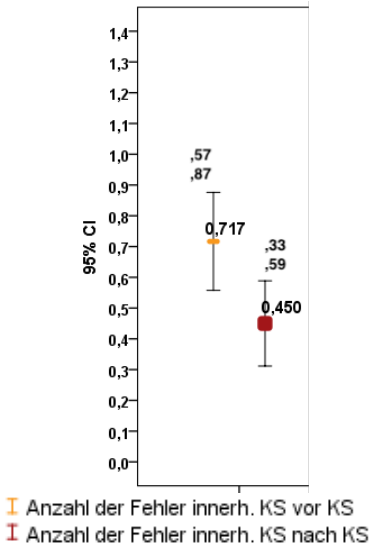
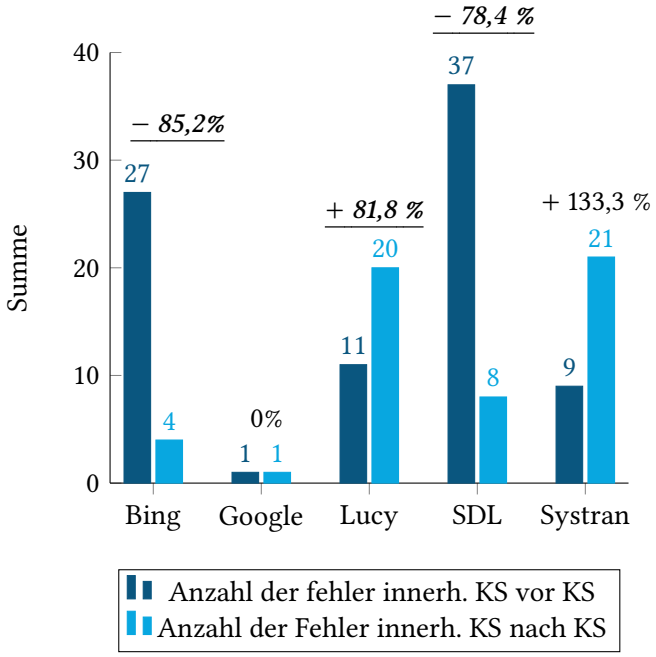


Abbildung 5.102: „Passiversatz verm.“ – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Tabelle 5.109: Beispiel 69

Vor-KS	Bei Funktionsstörungen **ist** die Maschine sofort **auszuschalten**.
HMÜ Bing	In the event of malfunction, the machine is immediately turn off.
Nach-KS	Bei Funktionsstörungen **schalten Sie** die Maschine sofort **aus**.
HMÜ Bing	In the event of malfunction, turn off the machine immediately.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



Signifikante Differenz vor vs. nach KS

Abbildung 5.103: „Passiversatz verm." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Bei dem NMÜ-System Google Translate wurden 23 der 24 analysierten Sätze sowohl vor als auch nach der Anwendung der KS-Regel korrekt übersetzt. Mit Ausnahme von Google waren die Differenzen bei allen anderen MÜ-Systemen signifikant (Abbildung 5.103): Die Abnahme bei Bing betrug 85,2 % (Mdiff = – ,958; z (N = 24) = – 3,573 / p < ,001) und bei SDL 78,4 % (Mdiff = – 1,208; z (N = 24) = – 3,815 / p < ,001). Auf der anderen Seite belief sich die Zunahme bei Lucy auf 81,8 % (Mdiff = ,375; z (N = 24) = – 2,496 / p = ,013) und bei Systran auf 133,3 % (Mdiff = ,500; z (N = 24) = – 2,126 / p = ,033). In Tabelle 5.110 werden die MÜ von SDL vs. Google verglichen:

Die Passiversatzkonstruktion war für SDL problematisch zu übersetzen. SDL hatte auf semantischer Ebene (die Ambiguität des Verbs ‚belegen' und auf syntaktischer Ebene (die Passiversatzkonstruktion) Schwierigkeiten bei der Übersetzung des Satzes vor und nach der Regelanwendung. Die Regelanwendung erleichterte für SDL die Übersetzung auf syntaktischer Ebene, dennoch blieb das Problem der Ambiguität des Verbs ungelöst. Auf der anderen Seite war Google in der Lage, den Satz vor und nach der Regelanwendung fehlerfrei zu übersetzen.

Tabelle 5.110: Beispiel 70

Vor-KS	Das Kaufdatum **ist** durch eine Kaufquittung **zu belegen**.
SMÜ SDL	The purchase date is through a purchase receipt to prove.
GNMÜ	The purchase date is to be confirmed by a purchase receipt.
Nach-KS	**Belegen Sie** das Kaufdatum durch eine Kaufquittung.
SMÜ SDL	Assign the purchase date by a purchase receipt.
GNMÜ	Confirm the purchase date with a purchase receipt.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.8.3 Aufteilung der Annotationsgruppen

Die größte Annotationsgruppe bei dieser Regel war die Gruppe RR; etwa 39 % der analysierten Sätze wurden sowohl mit als auch ohne die Passiversatzkonstruktion fehlerfrei übersetzt (Abbildung 5.104). Eine weitere große Gruppe war FR, in der knapp 28 % der Sätze erst nach der Regelanwendung korrekt übersetzt werden konnten. Die Gruppe FF war ebenfalls relativ groß; knapp 21 % der Sätze konnten weder vor noch nach der Regelanwendung korrekt übersetzt werden.

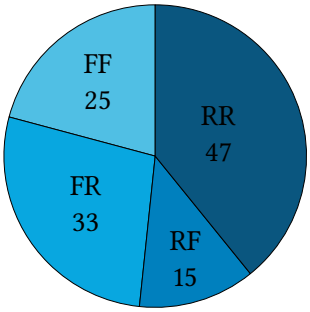
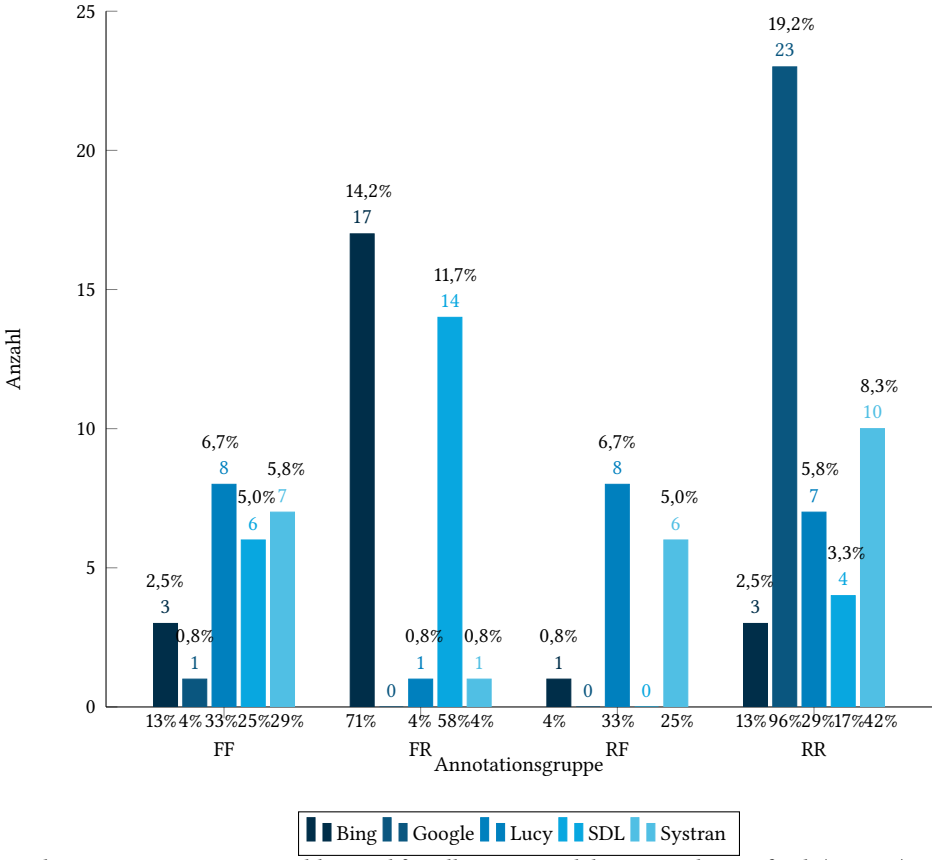


Abbildung 5.104: „Passiversatz verm.“ – Aufteilung der Annotationsgruppen

Die kleinste Gruppe war RF (ca. 12 %), in der die Passiversatzkonstruktion fehlerfrei übersetzt wurde und erst nach der Regelanwendung beinhaltete die Übersetzung Fehler (Abbildung 5.104). Im kommenden Abschnitt (§5.4.8.4) werden die aufgetretenen Fehlertypen analysiert.

5.4.8.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Eine Analyse der Annotationsgruppenaufteilung zeigte Folgendes (Abbildung 5.105):



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.105: „Passiversatz verm.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen

Bei zwei Systemen wurde die Mehrheit der falschen Übersetzungen mit Passiversatz (vor KS) nach der Regelanwendung korrekt übersetzt (Gruppe FR). Diese Systeme sind das HMÜ-System Bing mit 71 % und das SMÜ-System SDL mit 58 %. Auch bei dieser Regel waren 96 % der Übersetzungen des NMÜ-Systems Google

mit und ohne die Verwendung eines Passiversatzes korrekt (Gruppe RR). Dieser Prozentsatz ist mehr als doppelt so hoch wie die RR-Gruppe bei dem HMÜ-System Systran (42 %) und mehr als dreimal so hoch wie bei dem RBMÜ-System Lucy (29 %).

5.4.8.4 Vergleich der Fehlertypen mit vs. ohne die Konstruktion „sein + zu + Infinitiv“

Bei dieser Regel gab es signifikante Differenzen in der Fehleranzahl bei vier Fehlertypen: einen Anstieg beim lexikalischen Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ sowie einen Rückgang bei den grammatischen Fehlertypen GR.8 „Falsches Verb (Zeitform, Komposition, Person)“, GR.9 „Kongruenzfehler“ und GR.10 „Falsche Wortstellung“. Der Unterschied bei den Fehlertypen LX.4, GR.8, GR.9 und GR.10 erwies sich wie folgt als signifikant: $p < ,001$ bei den Fehlertypen LX.4 und GR.8 bzw. $p = ,008$ bei den Fehlertypen GR.9 und GR.10 / N = 120 (Abbildung 5.106).

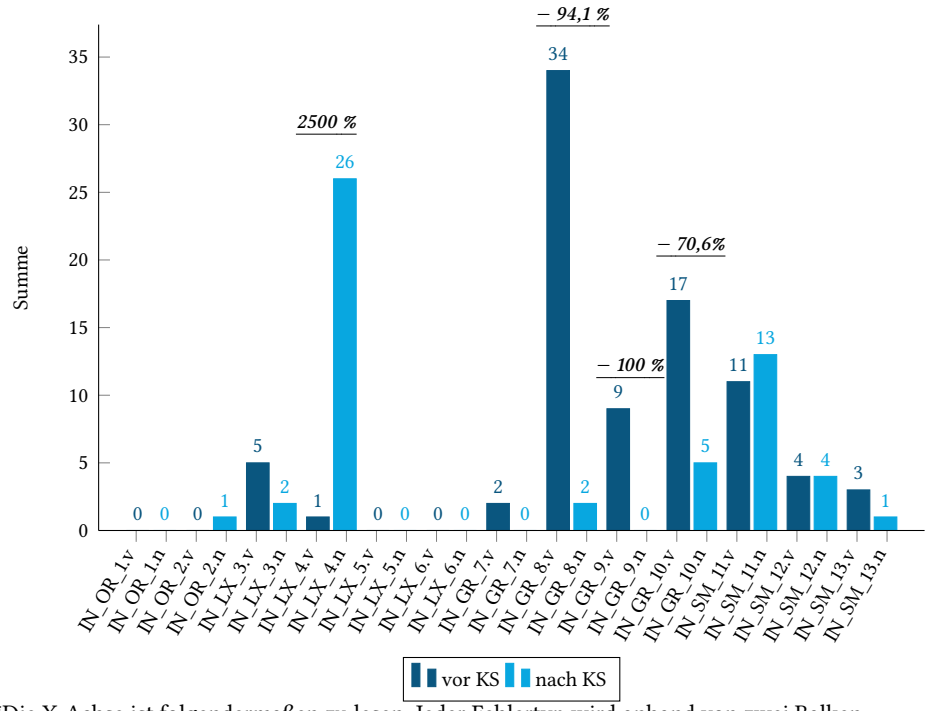
Um diese KS-Regel umzusetzen, wird der Passiversatz durch einen Imperativ ersetzt. Der Anstieg bei Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ von 1 auf 26 (2500 % / Mv = ,01 / SDv = ,091 / Mn = ,22 / SDn = ,414 / N = 120) entstand dadurch, dass einige Systeme den Imperativ (nach KS) nicht als solchen identifizieren konnten und stattdessen als normalen Ausgangssatz mit Subjekt ‚you‘ plus Verb übersetzten. Tabelle 5.111 demonstriert diesen Fall:

Tabelle 5.111: Beispiel 71

Vor-KS	Ist nur ein Gerät angeschlossen, so **ist** die Funktion PP **zu wählen**.
HMÜ Systran	If only one device is connected, the "PP" function is to be selected.
Nach-KS	Ist nur ein Gerät angeschlossen, so **wählen Sie** die Funktion PP.
HMÜ Systran	If only one device is connected, you select the "PP" function.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Hingegen sank die Fehleranzahl bei den Fehlertypen GR.8 „Falsches Verb (Zeitform, Komposition, Person)“ von 34 auf 2 (– 94,1 % / Mv = ,28 / SDv = ,453 / Mn = ,02 / SDn = ,129 / N = 120) und GR.10 „Falsche Wortstellung“ von 17 auf 5 (– 70,6 %



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler vor KS und der zweite die Summe der Fehler nach KS, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
__*Signifikante Differenz vor vs. nach KS***__
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.106: „Passiversatz verm.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

/ Mv = ,14 / SDv = ,350 / Mn = ,04 / SDn = ,201 / N = 120). Zudem wurde der Fehlertyp GR.9 „Kongruenzfehler (Agreement)“ nach der Regelanwendung vollständig behoben, von 9 auf 0, (− 100 % / Mv = ,08 / SDv = ,264 / Mn = − / SDn = − / N = 120). Die grammatische Konstruktion eines Passiversatzes „sein + zu + Infinitiv“ erschwert die syntaktische Analyse (Parsing) und lässt die MÜ ein falsches Verb, falsche Kongruenz bzw. falsche Wortstellung produzieren. Wenn wir den vorherigen Beispielsatz bei dem System Bing genauer betrachten, leuchtet diese grammatische Schwierigkeit ein (Tabelle 5.112).

Tabelle 5.112: Beispiel 72

Vor-KS	Ist nur ein Gerät angeschlossen, so **ist** die Funktion "PP" **zu wählen**.
HMÜ Bing	If only one device is connected, the function is "PP" to choose.
Nach-KS	Ist nur ein Gerät angeschlossen, so **wählen Sie** die Funktion "PP".
HMÜ Bing	If only one device is connected, select the "PP" function.

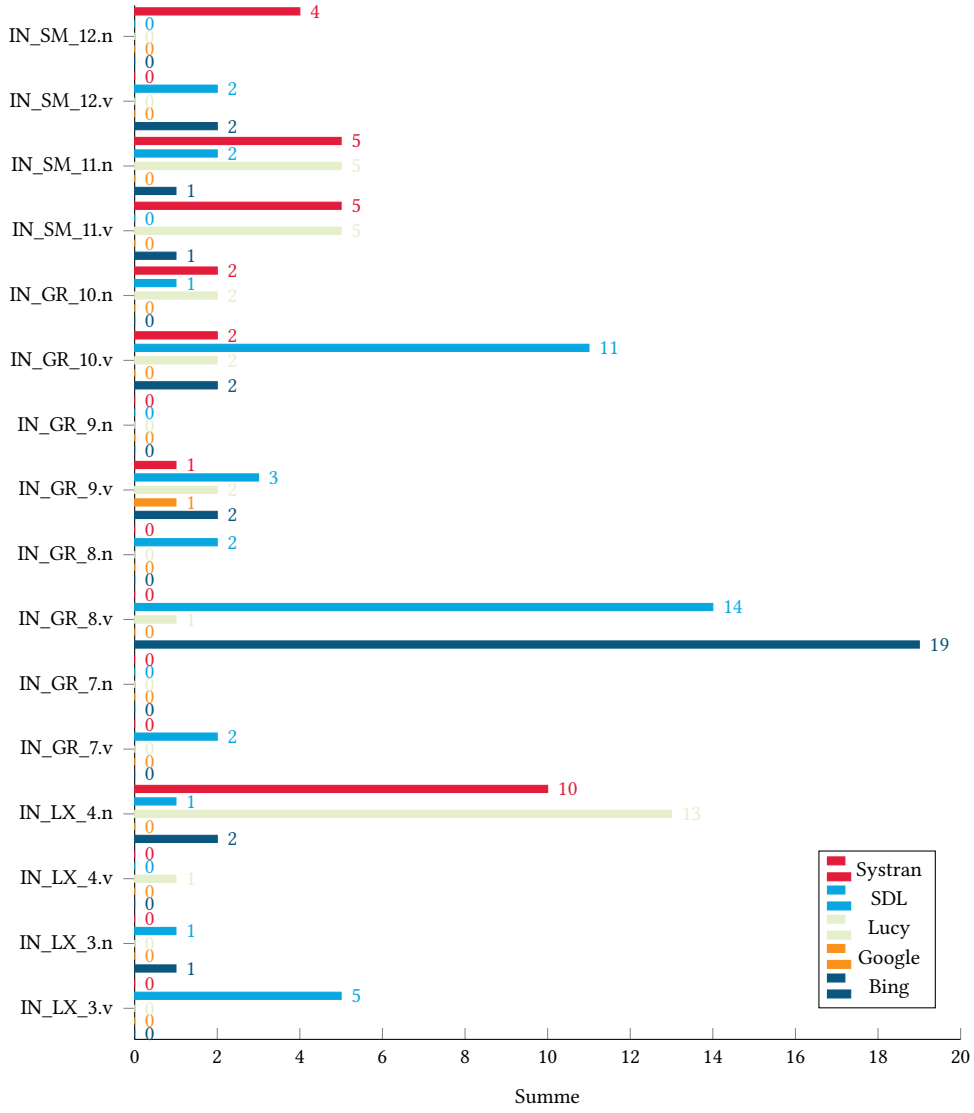
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Die MÜ der Passiversatzkonstruktion wurde falsch gebildet (GR.8) und platziert (GR.10). In diesem Fall unterstützte die Regel Bing bei der Eliminierung beider Fehlertypen.

5.4.8.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt (Abbildung 5.107), dass ein signifikanter Unterschied in der Fehleranzahl des Fehlertyps LX.4 „Zusätzliches Wort eingefügt“ bei dem RBMÜ-System Lucy von 1 auf 13 (+ 1200 %) und dem HMÜ-System Systran von 0 auf 10 (+ 100 %) zu beobachten ist (Tabelle 5.113).

Aus den signifikant veränderten grammatischen Fehlertypen GR.8, GR.9 und GR.10 wurde der Fehlertyp GR.8 „Falsches Verb (Zeitform, Komposition, Person)“ bei dem HMÜ-System Bing (von 19 auf 0; − 100 %) nach der Anwendung der KS-Regel vollständig behoben und sank deutlich bei dem SMÜ-System SDL (von 14 auf 2; − 85,7 %). Der Fehlertyp GR.9 „Kongruenzfehler (Agreement)“ zeigte bei keinem bestimmten System eine signifikante Veränderung. Der Fehlertyp GR.10 „Falsche Wortstellung“ verringerte sich bei dem SMÜ-System SDL (von 11 auf 1; − 90,9 %). Die Differenz in der Fehleranzahl bei den Fehlertypen LX.4, GR.8 und



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik traten die Fehlertypen 1, 5 und 6 bei gar keinem MÜ-System auf. Zudem wurden die Fehlertypen 2 und 13 nur einmal jeweils bei 1–3 MÜ-Systemen in vereinzelten Fällen registriert.

Abbildung 5.107: „Passiversatz verm.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

GR.10 erwies sich in den genannten Systemen folgendermaßen als signifikant (Tabelle 5.113).

Tabelle 5.113: „Passiversatz verm." – Fehlertypen mit signifikanter Veränderung nach KS

	N	Mittelwert	Standard-abweichung	Signifikanz (McNemar-Test)
LX.4 „Zusätzliches Wort eingefügt"				
Lucy	24	vor KS = ,04	vor KS = ,204	p < ,001
		nach KS = ,54	nach KS = ,509	
Systran	24	vor KS = –	vor KS = –	p = ,004
		nach KS = ,42	nach KS = ,504	
GR.8 „Falsches Verb (Zeitform, Komposition, Person)"				
Bing	24	vor KS = ,79	vor KS = ,415	p < ,001
		nach KS = –	nach KS = –	
SDL	24	vor KS = ,58	vor KS = ,504	p < ,001
		nach KS = ,08	nach KS = ,282	
GR.10 „Falsche Wortstellung"				
SDL	24	vor KS = ,46	vor KS = ,509	p = ,002
		nach KS = ,04	nach KS = ,204	

Die Verwendung eines Imperativs anstelle eines Passiversatzes vereinfachte die Satzstruktur für die älteren MÜ-Ansätze. Bei dem Neuronalen Ansatz hingegen konnte Google mit Ausnahme von einem Satz alle Sätze sowohl mit dem Passiversatz (vor KS) als auch mit dem Imperativ (nach KS) korrekt übersetzen (siehe Tabelle 5.110).

5.4.8.5 Vergleich der MÜ-Qualität mit vs. ohne die Konstruktion „sein + zu + Infinitiv" sowie die Korrelation zwischen den Fehlertypen und der Qualität

Sowohl die Stil- als auch die Inhaltsqualität[52] stiegen nach der Verwendung des Imperativs anstelle der Passiversatzkonstruktion deutlich an (Abbildung 5.108): Die Stilqualität nahm um 12,9 % zu (Mv = 3,73 / SDv = ,666 / Mn = 4,21 / SDn = ,569 / N = 97). Die Inhaltsqualität erhöhte sich um 10,6 % (Mv = 4,17 / SDv = ,886

[52]Definitionen der Qualität unter §4.5.5.1.

/ Mn = 4,61 / SDn = ,633 / N = 97). Der Mittelwert der Differenz (nach KS – vor KS) der vergebenen Qualitätspunkte pro Satz lag für die Stilqualität bei ,474 (SD = ,697) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,334 und einem Maximum von ,615 und für die Inhaltsqualität bei ,434 (SD = ,999) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,233 und einem Maximum von ,636 (Bootstrapping mit 1000 Stichproben) (Abbildung 5.109). Die Differenzen (nach KS – vor KS) in der Stil- und Inhaltsqualität erwiesen sich als hochsignifikant (z (N = 97) = – 5,758 / p < ,001) bzw. (z (N = 97) = – 3,806 / p < ,001).

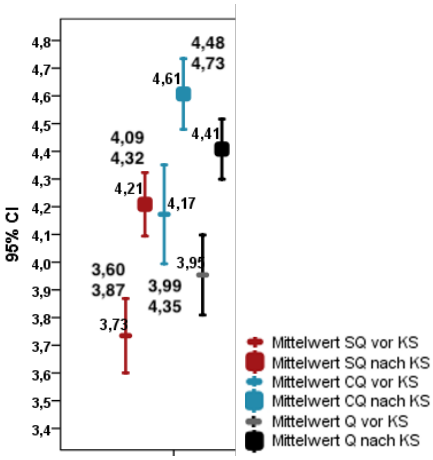


Abbildung 5.108: „Passiversatz verm." – Mittelwerte der Qualität vor und nach KS

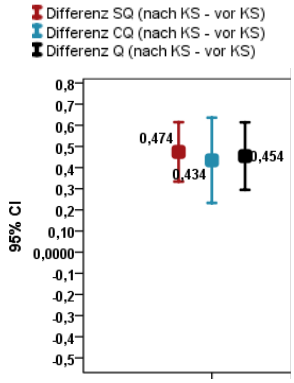
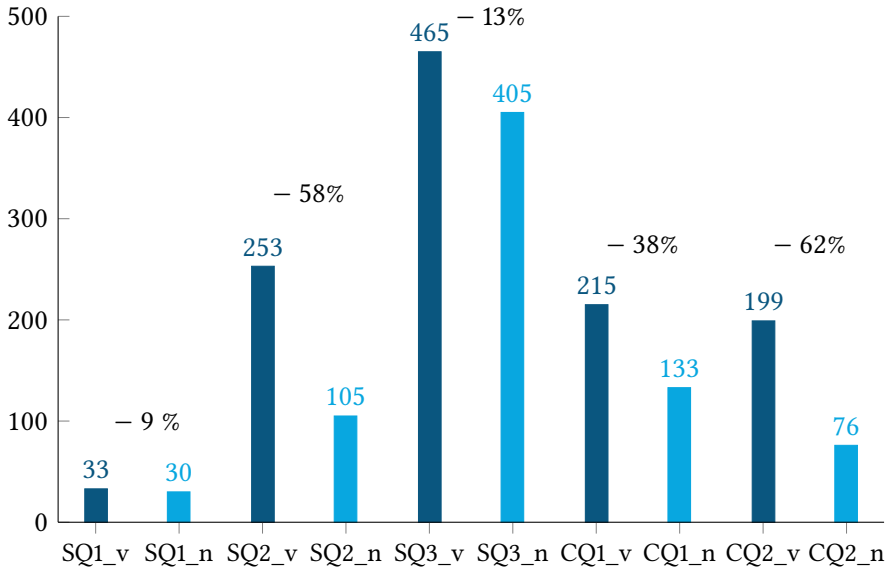


Abbildung 5.109: „Passiversatz verm." – Mittelwert der Qualitätsdifferenzen

Die Humanevaluation deckt den Grund des Qualitätsanstiegs auf. Wie Abbildung 5.110 zeigt, spielten hierbei das zweite Stilqualitätskriterium (SQ2) zusammen mit dem ersten und zweiten Inhaltsqualitätskriterium (CQ1 und CQ2) die wesentliche Rolle bei der Qualitätsveränderung.

Tabelle 5.114 verdeutlicht, wie die genannten Qualitätskriterien durch die Regelanwendung beeinflusst wurden.

In diesem Satz wurde der Nutzer (nach KS) durch den Imperativ direkt angesprochen. Dies erhöhte laut der Humanevaluation die Motivierung zum Handeln (SQ2). Außerdem wurden durch die Vereinfachung der Satzstruktur grammatische Fehler wie die falsche Verbform und Wortstellung korrigiert. Dadurch zeigte sich in der Humanevaluation eine deutliche Steigerung der Genauigkeit und Verständlichkeit der Übersetzung (CQ1 und CQ2).



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder.
CQ2: Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.110: „Passiversatz verm.“ – Vergleich der Qualitätskriterien

Tabelle 5.114: Beispiel 73

Vor-KS	Ist nur ein Gerät angeschlossen, so **ist** die Funktion PP **zu wählen**.
HMÜ Bing	If only one device is connected, the function is "PP" to choose.
Nach-KS	Ist nur ein Gerät angeschlossen, so **wählen Sie** die Funktion PP.
HMÜ Bing	If only one device is connected, select the "PP" function.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.8.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung in der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS – Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS – Q. vor KS) zusammenhängt. Mithilfe des Spearman-Tests erwies sich ein signifikanter starker negativer Zusammenhang zwischen der Differenz im Fehlertyp GR.8 „Falsches Verb (Zeitform, Komposition, Person)“ und der Differenz in der Stilqualität. Außerdem erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.4 „Zusätzliches Wort eingefügt“ und GR.10 „Wortstellungsfehler“ einzeln und der Differenz in der Stilqualität; sowie ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.3 „Wort ausgelassen“, GR.7 „Falsche Wortart / Wortklasse“ und SM.12 „Falsche Wahl“ einzeln und der Differenz in der Stilqualität. (siehe Tabelle 5.115)

Bezüglich der Inhaltsqualität erwies sich ein signifikanter starker negativer Zusammenhang zwischen der Differenz im Fehlertyp GR.8 „Falsches Verb (Zeitform, Komposition, Person)“ und der Differenz in der Inhaltsqualität; ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz im Fehlertyp GR.10 „Wortstellungsfehler“ und der Differenz in der Inhaltsqualität; sowie ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.3 „Wort ausgelassen“, LX.4 „Zusätzliches Wort eingefügt“, GR.7 „Falsche Wortart / Wortklasse“ und SM.12 „Falsche Wahl“ einzeln und der Differenz in der Inhaltsqualität. (siehe Tabelle 5.115)

Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht erwiesen werden.

Diese signifikanten negativen Korrelationen deuten darauf hin, dass mit dem Rückgang der Fehleranzahl der genannten Fehlertypen nach KS ein Qualitätsanstieg verzeichnet wurde. In Tabelle 5.114 wurden die grammatischen Fehlertypen GR.8 „Falsches Verb“ und GR.10 „Wortstellung“ in ‚is … to choose‘ eliminiert, daraufhin stieg die Stilqualität um 1,63 Punkte und die Inhaltsqualität um 1,75 Punkte auf der Likert-Skala. In Tabelle 5.116 wurde nur der Verbformfehler (GR.8) in ‚replace‘ korrigiert, nachdem der Imperativ anstelle des Passiversatzes verwendet wurde.

Daraufhin stiegen die Stilqualität um 1,00 Punkte und die Inhaltsqualität um 0,75 Punkte auf der Likert-Skala.

Tabelle 5.115: „Passiversatz verm.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	97	,023	− ,230
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	97	< ,001	− ,448
Diff. der Anzahl der **GR.7 „Falsche Wortart“**	97	,024	− ,229
Diff. der Anzahl der **GR.8 „Falsches Verb“**	97	< ,001	− ,617
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	97	< ,001	− ,438
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	97	,005	− ,282
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	97	,030	− ,220
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	97	,005	− ,286
Diff. der Anzahl der **GR.7 „Falsche Wortart“**	97	,014	− ,248
Diff. der Anzahl der **GR.8 „Falsches Verb“**	97	< ,001	− ,641
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	97	,001	− ,324
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	97	,008	− ,267
Differenz allg. Q (nach KS − vor KS)			
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	97	,031	− ,219
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	97	< ,001	− ,367
Diff. der Anzahl der **GR.7 „Falsche Wortart“**	97	,016	− ,244
Diff. der Anzahl der **GR.8 „Falsches Verb“**	97	< ,001	− ,681
Diff. der Anzahl der **GR.10 „Wortstellungsfehler“**	97	< ,001	− ,418
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	97	,004	− ,287

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Tabelle 5.116: Beispiel 74

Vor-KS	Die Bedienungsanleitung **ist** im Falle des Verlustes **zu ersetzen**.
HMÜ Bing	The operating instructions are to replace in case of loss.
Nach-KS	**Ersetzen Sie** die Bedienungsanleitung im Falle des Verlustes.
HMÜ Bing	Replace the operating instructions in case of loss.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.8.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.111 zeigt, nahmen nicht bei allen Systemen die Stil- und Inhaltsqualität nach der Anwendung der KS-Regel zu.

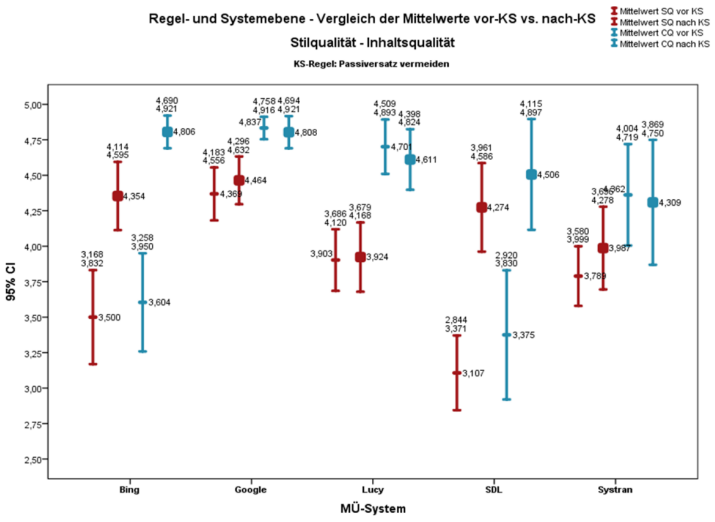


Abbildung 5.111: „Passiversatz verm.“ – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Die Stilqualität stieg bei dem SMÜ-System SDL (+ 37,6 %), dem HMÜ-System Bing (+ 24,4 %) sowie dem HMÜ-System Systran (+ 5,2 %) signifikant, während die Inhaltsqualität nur bei Bing (+ 33,4 %) und SDL (+ 33,5 %) signifikant zunahm (Tabelle 5.117).

Tabelle 5.117: „Passiversatz verm.“ – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	18	,001	− 3,342	18	< ,001	− 3,519	18	< ,001	− 3,616
Google	21	,104	− 1,624	21	,774	− ,287	21	,229	− 1,203
Lucy	18	,819	− ,229	18	,205	− 1,268	18	,498	− ,678
SDL	21	< ,001	− 4,018	21	,003	− 2,958	21	< ,001	− 3,702
Systran	19	,041	− 2,045	19	,551	− ,597	19	,034	− 2,118

p: Signifikanz z: Teststatistik nicht signifikant ($p \geq 0{,}05$)

Wie die Aufteilung der Annotationsgruppen zeigte, waren 96 % der Übersetzungen (23 von 24 Sätzen) des NMÜ-Systems Google Translate sowohl vor als auch nach der Regelanwendung (Annotationsgruppe RR) richtig. Entsprechend ist zu erwarten, dass das Qualitätsniveau sich nicht verändert.

Bei dem RBMÜ-System Lucy waren die Ergebnisse der Annotationsgruppen (Abbildung 5.105) sehr gemischt (d. h. die Zahlen der Gruppen FF, RF und RR fielen ähnlich hoch aus), daher konnte keine signifikante Veränderung in der Qualität nachgewiesen werden.

5.4.8.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwies sich bei dem HMÜ-System Bing ein signifikanter starker negativer Zusammenhang zwischen den grammatischen Fehlertypen GR.8 „Falsches Verb (Zeitform, Komposition, Person)“ und GR.10 „Wortstellungsfehler“ einzeln und der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen GR.8 und der Inhaltsqualität (Tabelle 5.118).

Bei dem RBMÜ-System Lucy erwies sich nur eine signifikante starke negative Korrelation zwischen Fehlertyp GR.8 „Falsches Verb“ und der Inhaltsqualität (Tabelle 5.118).

Bei dem SMÜ-System SDL erwies sich nur eine signifikante starke negative Korrelation zwischen Fehlertyp GR.7 „Falsche Wortart / Wortklasse“ und der Inhaltsqualität (Tabelle 5.118).

Bei dem HMÜ-System Systran erwies sich eine signifikante starke negative Korrelation zwischen dem lexikalischen Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ und der Stilqualität sowie eine signifikante starke negative Korrelation zwischen den Fehlertypen LX.4 „Zusätzliches Wort eingefügt“ und GR.10 „Wortstellungsfehler“ einzeln und der Inhaltsqualität (Tabelle 5.118).

5.4.8.6 Vergleich der MÜ-Qualität mit vs. ohne die Konstruktion „sein + zu + Infinitiv“ auf Annotationsgruppenebene

Die Stil- und Inhaltsqualität[53] der MÜ stiegen in den Annotationsgruppen FF und FR. In der Annotationsgruppe RR stieg nur die Stilqualität, während die Inhaltsqualität unverändert blieb. In der Annotationsgruppe RF sanken beide Qualitätswerte. (siehe Abbildung 5.112)

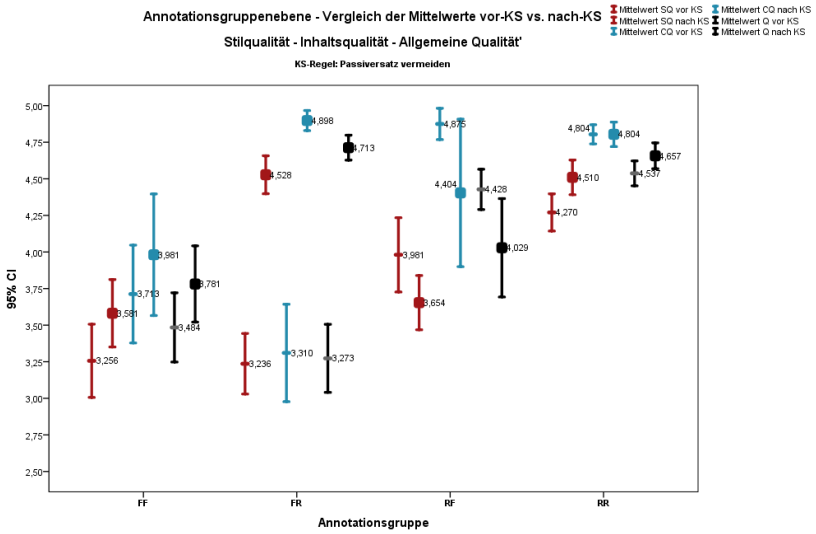


Abbildung 5.112: „Passiversatz verm.“ – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

Nicht alle genannten Qualitätsveränderungen waren statistisch signifikant. Tabelle 5.119 zeigt die Signifikanz der Veränderungen der Qualitätswerte jeder Annotationsgruppe.

[53] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.118: „Passiversatz verm.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Bing			Lucy			SDL			Systran		
	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ
Differenz SQ (nach KS − vor KS)												
Diff. der Anzahl **LX.4 „W. extra eingefügt“**										19	,006	− ,605
Diff. der Anzahl **GR.7 „Falsche Wortart“**							21	,060	− ,417			
Diff. der Anzahl **GR.8 „Falsches Verb“**	18	,006	− ,620	18	,051	− ,466						
Diff. der Anzahl **GR.10 „Wortstellungsfeh.“**	18	,029	− ,513							19	,608	− ,126
Differenz CQ (nach KS − vor KS)												
Diff. der Anzahl **LX.4 „W. extra eingefügt“**										19	,006	− ,605
Diff. der Anzahl **GR.7 „Falsche Wortart“**							21	,018	− ,511			
Diff. der Anzahl **GR.8 „Falsches Verb“**	18	,008	− ,605	18	,003	− ,656						
Diff. der Anzahl **GR.10 „Wortstellungsfeh.“**	18	,054	− ,461							19	,004	− ,634
Differenz Q (nach KS − vor KS)												
Diff. der Anzahl **LX.4 „W. extra eingefügt“**										19	,539	− ,150
Diff. der Anzahl **GR.7 „Falsche Wortart“**							21	,018	− ,509			
Diff. der Anzahl **GR.8 „Falsches Verb“**	18	,004	− ,647	18	,007	− ,608						
Diff. der Anzahl **GR.10 „Wortstellungsfeh.“**	18	,030	− ,512							19	,007	− ,601

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (ρ ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Tabelle 5.119: „Passiversatz verm.“ – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	20	,003	− 3,021
Differenz CQ (nach KS − vor KS)	20	,136	− 1,492
Differenz allg. Q (nach KS − vor KS)	20	,012	− 2,505
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	27	< ,001	− 4,463
Differenz CQ (nach KS − vor KS)	27	< ,001	− 4,544
Differenz allg. Q (nach KS − vor KS)	27	< ,001	− 4,542
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	13	,016	− 2,417
Differenz CQ (nach KS − vor KS)	13	,007	− 2,680
Differenz allg. Q (nach KS − vor KS)	13	,001	− 3,192
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	37	< ,001	− 3,737
Differenz CQ (nach KS − vor KS)	37	,522	− ,640
Differenz allg. Q (nach KS − vor KS)	37	,002	− 3,167

In der Gruppe FF (Übersetzung mit und ohne Passiversatz falsch) stieg die Stilqualität signifikant (z (N = 20) = − 3,021 / p = ,003) und die Inhaltsqualität nicht signifikant (z (N = 20) = − 1,492 / p = ,136). Der Qualitätsanstieg in dieser Gruppe zeigt, dass die Bewerter die Fehler bei der Verwendung eines Passiversatzes (vor KS) schwerwiegender als die bei der Vermeidung eines Passiversatzes (nach KS) aufgetretenen Fehler wahrnahmen. Dieses Ergebnis wird außerdem durch den Analysefaktor 4 „Korrelation zwischen den Fehlertypen und der Qualität“ (§5.4.8.5) ersichtlich. Tabelle 5.120 verdeutlicht diesen unterschiedlichen Schweregrad der aufgetretenen Fehler in den beiden Szenarien, vor vs. nach der Regelanwendung.

Bei der Verwendung eines Passiversatzes (vor KS) trat der grammatische Fehler GR.8 „Falsches Verb“ (in ‚are to lock‘) auf. Die Bewerter fanden die Übersetzung irreführend und begründeten dies wie folgt: „'are to lock' sounds like the

Tabelle 5.120: Beispiel 75

Vor-KS	Ist ein mehrstufiges Modul parametriert, so **sind** die externen Kontakte **zu verriegeln**.
HMÜ Bing	If a multi-stage module is parameterized, the external contacts are to lock.
Nach-KS	Ist ein mehrstufiges Modul parametriert, **verriegeln Sie** die externen Kontakte.
HMÜ Bing	If a multi-stage module is parameterized, you lock the external contacts.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

contacts will lock themselves. I suggest 'are to be locked' instead". Ein anderer Bewerter empfahl die Verwendung des Imperativs und kommentierte: „The text does not make clear that the reader should act at all". Bei der Verwendung eines Imperativs (nach KS) trat der lexikalische Fehler LX.4 „Zusätzliches Wort eingefügt" (in ‚you') auf. Die Bewerterkommentare zeigen, dass der Imperativ durch diesen Fehler nicht klar war und somit der Stil den Leser nicht adäquat zum Handeln anregen konnte. Ein Bewerterkommentar lautete „'you' should be deleted to keep the sentence in the imperative and motivate the user to act". Durch die vergebenen Scores in den beiden Szenarien wird ersichtlich, dass die Bewerter den Fehlertyp im vor-KS-Szenario schwerwiegender als im nach-KS-Szenario empfanden, denn die Stil- und Inhaltsqualität stiegen nach KS um + 0,25 Punkte bzw. + 1,50 Punkte auf der Likert-Skala.

Erwartungsgemäß nahmen die Stil- und Inhaltsqualität in der Gruppe FR (MÜ falsch vor KS; richtig nach KS) zu und sanken in der Gruppe RF (MÜ richtig vor KS; falsch nach KS) signifikant (siehe Tabelle 5.119), wobei die Qualitätsdifferenz in der Gruppe FR deutlich höher (Diff_SQ + 39,9 %; Diff_CQ + 48,0 %) als in der Gruppe RF ausfiel (Diff_SQ − 8,2 %; Diff_CQ − 9,7 %). Gleichzeitig zeigte die Aufteilung der Annotationsgruppen (siehe §5.4.8.3), dass die Gruppe FR (28 %) mehr als doppelt so häufig vertreten war als die Gruppe RF (13 %), siehe Abbildung 5.104. Auf Basis dieses Vergleichs der Gruppen FR und RF lässt sich schlussfolgern, dass der potenzielle Qualitätsgewinn bei der Vermeidung eines Passiversatzes (Anwendung der KS-Regel) sehr hoch ausfällt.

In der Gruppe RR (Übersetzung mit und ohne Passiversatz richtig) stieg die Stilqualität signifikant ($z (N = 37) = -3{,}737 / p < ,001$), während die Inhaltsqualität

unverändert blieb. Dieses Ergebnis ist nachvollziehbar, denn dank einer richtigen Übersetzung eines Passiversatzes (vor KS) oder eines Imperativs (nach KS) sind beide inhaltlich verständlich und präzise. Aus stilistischer Sicht fanden die Bewerter jedoch die Verwendung des Passiversatzes weniger geeignet für die Satzintention. Tabelle 5.121 ist für einen Beispielsatz aus der Gruppe RR. Hierbei fanden die Bewerter den Stil bei der Verwendung des Imperativs (nach KS) geeigneter. Entsprechend stieg die Stilqualität um + 0,50 Punkte auf der Likert-Skala, während die Inhaltsqualität unverändert blieb.

Tabelle 5.121: Beispiel 76

Vor-KS	Das Kaufdatum **ist** durch eine Kaufquittung **zu belegen**.
GNMÜ	The purchase date is to be confirmed by a purchase receipt.
Nach-KS	**Belegen Sie** das Kaufdatum durch eine Kaufquittung.
GNMÜ	Confirm the purchase date with a purchase receipt.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

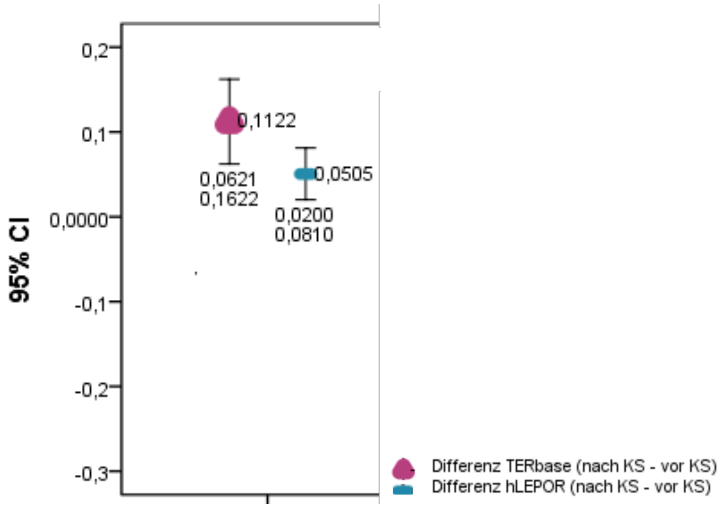
5.4.8.7 Vergleich der AEM-Scores mit vs. ohne die Konstruktion „sein + zu + Infinitiv“ sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores mit und ohne Verwendung des Passiversatzes zeigte sowohl mit TERbase als auch mit hLEPOR eine deutliche Verbesserung der AEM-Scores.

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei ,112 (SD = ,248) und für die hLEPOR bei ,051 (SD = ,152) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.113. Die Differenzen (nach KS − vor KS) in TERbase und hLEPOR erwiesen sich als signifikant (z (N = 97) = − 4,175 / p < ,001) bzw. (z (N = 97) = − 3,396 / p = ,001). Dieses Ergebnis deutet darauf hin, dass nach der Verwendung des Imperativs (nach KS) weniger Edits erforderlich waren.

5.4.8.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Nach der Anwendung der KS-Regel stieg die Qualität signifikant (siehe §5.4.8.5). Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter mittlerer



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.113: „Passiversatz verm." – Mittelwert der Differenz der AEM-Scores

positiver Zusammenhang zwischen den Differenzen der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität. Nach der Verwendung des Imperativs anstelle des Passiversatzes (nach KS) verbesserten sich die Scores der beiden AEM-Score-Metriken und die Qualität nahm zu.

5.4.8.8 Analyse der siebten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der siebten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (−) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[54]

[54] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Tabelle 5.122: „Passiversatz verm.“ – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS – vor KS)	97	,002	,311	mittlerer Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS – vor KS)	97	< ,001	,473	mittlerer Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Regel 7: Konstruktionen mit „sein + zu + Infinitiv“ vermeiden

Erster Analysefaktor: Vergleich der Fehleranzahl mit vs. ohne die Konstruktion „sein + zu + Infinitiv“

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Die Fehleranzahl sank signifikant, nachdem die Passiversatzkonstruktion als Imperativ umformuliert wurde.

Anz.F. (–)

Auf Regel- und MÜ-Systemebene:
Bei Bing und SDL sank die Fehleranzahl nach KS signifikant.

Bi (–)
SD (–)

Bei Lucy und Systran stieg die Fehleranzahl nach KS signifikant.	**Lu (+)** **Sy (+)**
Bei Google gab es nur einen Fehler bei jedem Szenario.	**Go (=)**

Zweiter Analysefaktor

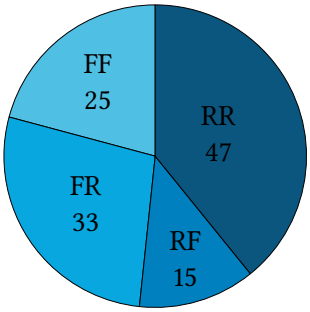


Abbildung 5.114: Aufteilung der Annotationsgruppen auf Regelebene

Dritter Analysefaktor: Vergleich der Fehlertypen mit vs. ohne die Konstruktion „sein + zu + Infinitiv“

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:

H1 wurde für vier Fehlertypen bestätigt. Die Fehleranzahl von LX.4 „Zusätzliches Wort eingefügt“ stieg signifikant.	**LX.4 (+)**
Die Fehleranzahl von GR.8 „Falsches Verb“, GR.9 „Kongruenzfehler“ und GR.10 „Wortstellungsfehler“ sanken nach KS signifikant.	**GR.8 (–)** **GR.9 (–)** **GR.10 (–)**

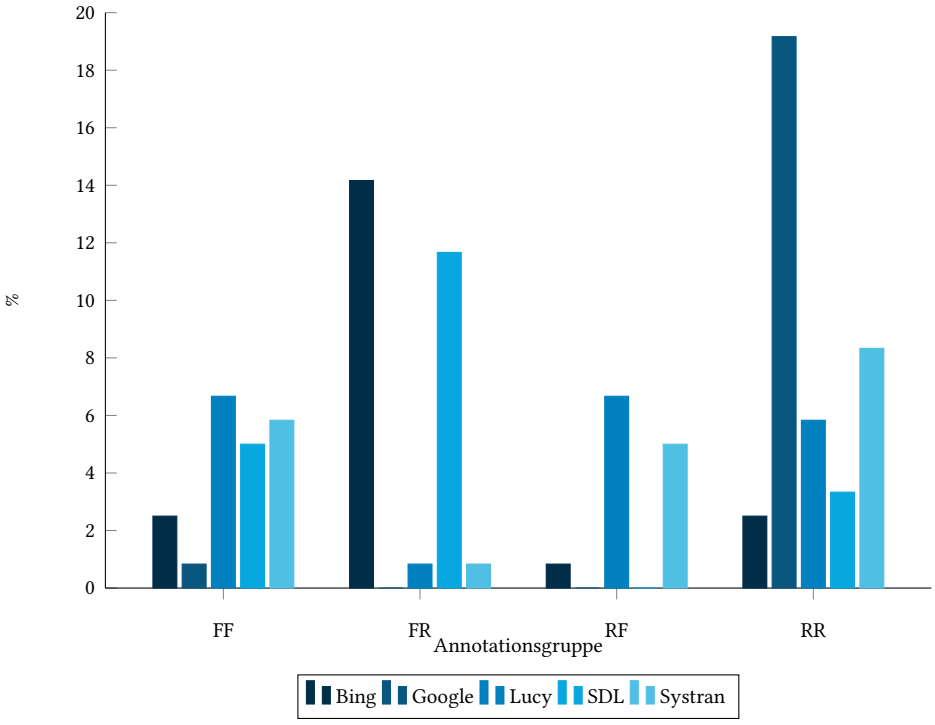


Abbildung 5.115: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Auf Regel- und MÜ-Systemebene:

Bei Lucy und Systran stieg die Fehleranzahl von LX.4 „Zusätzliches Wort eingefügt“ nach KS signifikant.	**LX.4 (+): Lu Sy**
Bei Bing und SDL sank die Fehleranzahl von GR.8 „Falsches Verb“ nach KS signifikant.	**GR.8 (−): Bi SD**
Bei SDL sank die Fehleranzahl von GR.10 „Wortstellungsfehler“ nach KS signifikant.	**GR.10 (−): SD**

Alle weiteren Veränderungen waren nicht signifikant.

Vierter Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne die Konstruktion „sein + zu + Infinitiv“

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Sowohl die Stil- als auch die Inhaltsqualität stiegen signifikant nach Verwendung des Imperativs (nach KS).

SQ (+)
CQ (+)

Auf Regel- und MÜ-Systemebene:
Die Stil- und Inhaltsqualität stiegen bei Bing, und SDL signifikant, wobei bei Systran nur die Stilqualität signifikant zunahm.

SQ (+):	**CQ (+):**
Bi SD	**Bi**
Sy	**SD**

Alle weiteren Qualitätsveränderungen waren nicht signifikant.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde für drei Fehlertypen bestätigt.
Es bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.8 „Falsches Verb“ und der Differenz der Stil- und Inhaltsqualität.

***neg* GR.8 <<>> SQ**
***neg* GR.8 <<>> CQ**

Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und der Differenz der Stilqualität sowie ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Wortstellungsfehler“ und der Differenz der Stil- und Inhaltsqualität.

***neg* LX.4 <> SQ**

***neg* GR.10 <> SQ**
***neg* GR.10 <> CQ**

Auf Regel- und MÜ-Systemebene:

Bei Bing bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.8 „Falsches Verb“ und der Differenz der Stil- und Inhaltsqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Wortstellungsfehler“ und der Differenz der Stilqualität.

Bi
***neg* GR.8 <<>> SQ**
***neg* GR.8 <<>> CQ**
***neg* GR.10 <<>> SQ**

Bei Lucy bestand ein signifikanter *starker* negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.8 „Falsches Verb“ und der Differenz der Inhaltsqualität.

Lu
***neg* GR.8 <<>> CQ**

Bei SDL bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.7 „Falsche Wortart“ und der Differenz der Inhaltsqualität.

SD
***neg* GR.7 <<>> CQ**

Bei Systran bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und der Differenz der Stil- und Inhaltsqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz der Fehleranzahl des GR.10 „Wortstellungsfehler“ und der Differenz der Inhaltsqualität.

Sy
***neg* LX.4 <<>> SQ**
***neg* LX.4 <<>> CQ**
***neg* GR.10 <<>> CQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne die Konstruktion „sein + zu + Infinitiv“ auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde nur zum Teil bestätigt: Bei der Annotationsgruppe FF stieg die Stilqualität signifikant und die Inhaltsqualität nicht signifikant.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe FR stiegen die Stil- und Inhaltsqualität signifikant.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe RF sanken die Stil- und Inhaltsqualität signifikant.	**SQ (−)** **CQ (−)**
Bei der Annotationsgruppe RR stieg die Stilqualität signifikant und die Inhaltsqualität blieb unverändert.	**SQ (+)** **CQ (=)**

Siebter Analysefaktor: Vergleich der AEM-Scores mit vs. ohne die Konstruktion „sein + zu + Infinitiv“ auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt. AEM-Scores sowohl von TERbase als auch von hLEPOR verbesserten sich nach Verwendung des Imperativs (nach KS) signifikant.	**TERbase (+)** **hLEPOR (+)**

Achter Analysefaktor: Korrelation zwischen den Differenzen der AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz der AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz der AEM-Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.9 ACHTE REGEL: Überflüssige Präfixe vermeiden

5.4.9.1 Überblick

Im Folgenden wird die KS-Regel „Überflüssige Präfixe vermeiden" kurz beschrieben.[55] Zudem wird zusammenfassend und anhand von Beispielen demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Überflüssige Präfixe vermeiden (tekom-Regel-Nr. L 114)

Nach dieser Regel sollen Verben mit Präfixen vermieden werden, wenn das Verb ohne Präfix die gleiche Bedeutung hat (tekom 2013: 111).

Begründung: Die Kürzung vereinfacht den Satz und reduziert Segmentvarianten bei der Übersetzung mithilfe eines Systems (ebd.).

Umsetzungsmuster:

Vor KS: Verb mit Präfix
Nach KS: Eliminierung des Präfixes

[55] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

KS-Stelle

Vor KS: Verb mit Präfix (trennbare und untrennbare Verben)
Nach KS: Verb ohne Präfix

Beispiele

Überprüfen *Sie, ob ausreichend Wasser im Wassertank vorhanden ist.*

Prüfen *Sie, ob ausreichend Wasser im Wassertank vorhanden ist.*

Wählen *Sie die Option "Software von einer bestimmten Liste installieren"* *aus* *.*

Wählen *Sie die Option "Software von einer bestimmten Liste installieren".*

Aufteilung der Testsätze: Die Platzierung des Präfixes bei trennbaren Verben kann den Schwierigkeitsgrad bzw. die Richtigkeit der Übersetzung beeinflussen, daher deckt der Datensatz Folgendes ab: 15 Sätze, in denen das Präfix getrennt am Satzende bzw. Nebensatzende erscheint sowie 9 Sätze mit untrennbaren Verben oder trennbaren Verben, in denen das Präfix ungetrennt vom Verb auftritt.

Im Folgenden werden die Ergebnisse der einzelnen Analysefaktoren dargestellt.

5.4.9.2 Vergleich der Fehleranzahl mit vs. ohne überflüssige Präfixe

Die Fehleranzahl sank deutlich um 35,5 % von 45 Fehlern bei der Verwendung von überflüssigen Präfixen (M = ,38 / SD = ,649 / N = 120) auf 29 Fehler bei der Vermeidung von überflüssigen Präfixen (M = ,24 / SD = ,550 / N = 120), siehe Abbildung 5.116. Der Mittelwert der Differenz (nach KS − vor KS) der Fehleranzahl pro Satz lag somit bei − ,13 (SD = ,517) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,23 (SD = ,401) und einem Maximum von − ,04 (SD = ,637) (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.117. Die Differenz (nach KS − vor KS) der Fehleranzahl erwies sich als signifikant (z (N = 120) = − 2,717 / p = ,007).

Insbesondere trennbare Verben waren oft schwer zu parsen; abhängig von der Satzstruktur stehen die Präfixe in manchen Fällen am Ende des Satzes – weit entfernt vom Rest des Verbs. In solchen Fällen konnten mehrere Systeme den Satz zwar korrekt übersetzen, fügten aber am Satzende eine redundante Übersetzung

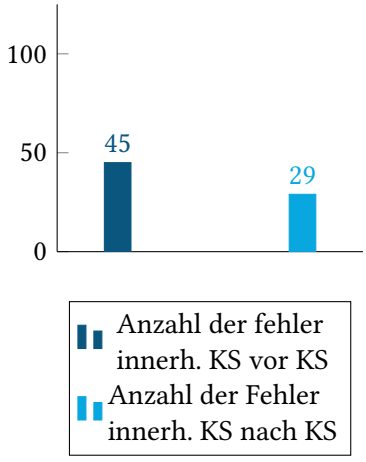


Abbildung 5.116: „Überfl. Präfixe verm.“ – Fehlersumme vor vs. nach KS

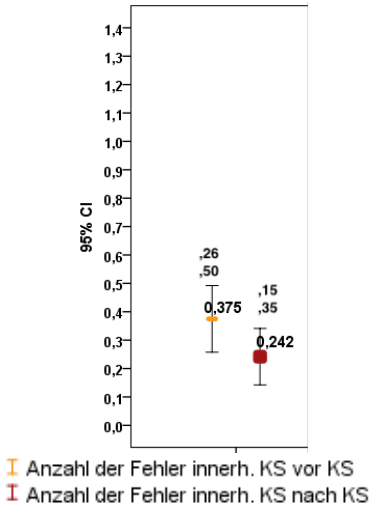


Abbildung 5.117: „Überfl. Präfixe verm.“ – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Tabelle 5.123: Beispiel 77

Vor-KS	**Schicken** Sie das Gerät zusammen mit dem Original-Kaufbeleg an nachstehende Adresse **zu**.
HMÜ Systran	Send the appliance together with the original receipt to the following address too.
Nach-KS	**Schicken** Sie das Gerät zusammen mit dem Original-Kaufbeleg an nachstehende Adresse.
HMÜ Systran	Send the appliance together with the original receipt to the following address.

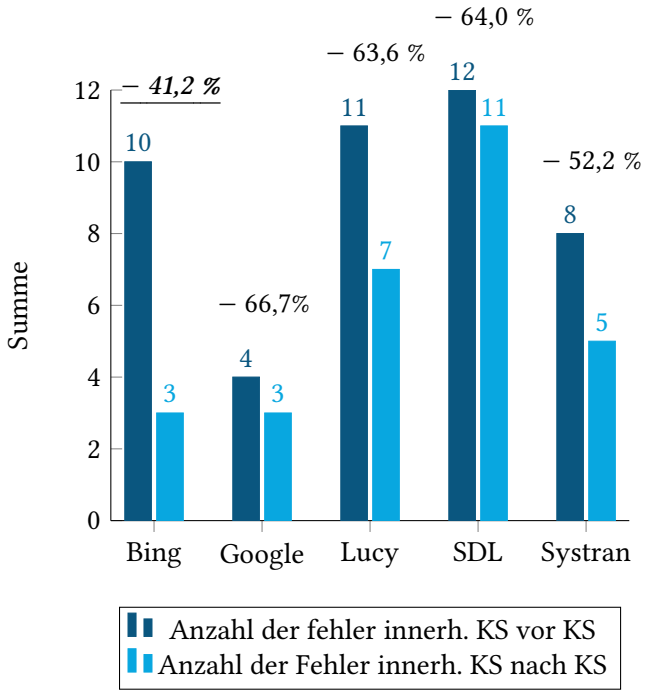
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

für das Präfix hinzu. In Tabelle 5.123 wurde ‚too' als Übersetzung von ‚zu' im Verb ‚zuschicken' zusätzlich hinzugefügt.

Daher war die Anwendung der Regel in dieser Hinsicht sinnvoll. Rund 59 % (14 von 24) der analysierten Sätze wurden vor der Regelanwendung von mindestens einem MÜ-System falsch übersetzt und mithilfe der Regel korrigiert.

5.4.9.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Die Fehleranzahl nach der Umsetzung der KS-Regel sank bei allen Systemen (Abbildung 5.118): Im Allgemeinen war die Fehleranzahl relativ klein im Vergleich zu den anderen Regeln.



Signifikante Differenz vor vs. nach KS

Abbildung 5.118: „Überfl. Präfixe verm." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Signifikant war nur der Rückgang bei dem HMÜ-System Bing Mdiff = - 0,292; z (N = 24) = – 2,070 / p = ,038), siehe Abbildung 5.118. Bei den restlichen Systemen betrug der Mittelwert der Differenz in der Fehleranzahl (nach KS – vor KS):

HMÜ-System Systran (Mdiff = – ,125); RBMÜ-System Lucy (Mdiff = – ,167); NMÜ-System Google Translate (Mdiff = – ,042); SMÜ-System SDL (Mdiff = – ,042). Im Vergleich zu Systran in Tabelle 5.123 konnten SDL und Google Translate beide Szenarien fehlerfrei und identisch (als ‚Check if there is sufficient water in the water tank') übersetzen. In den nächsten Abschnitten werden die aufgetretenen Fehlertypen sowie die Aufteilung der Annotationsgruppen näher erläutert.

5.4.9.3 Aufteilung der Annotationsgruppen

Die größte Annotationsgruppe stellte die Gruppe RR dar; die MÜ-Systeme konnten mehr als zwei Drittel der Sätze sowohl mit als auch ohne die überflüssigen Präfixe fehlerfrei übersetzen (Abbildung 5.119). An der zweiten Stelle kommt die Gruppe FF, die relativ klein war (ca. 16 %). Die Gruppe FR (ca. 13 %) zeigt, dass die Regel die MÜ-Systeme dabei unterstützen konnte, aufgetretene Fehler zu beheben. Zum Schluss folgt die Gruppe RF mit einem sehr kleinen Anteil von ca. 3 % (Abbildung 5.119).

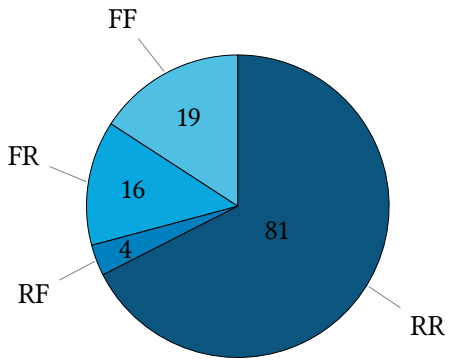


Abbildung 5.119: „Überfl. Präfixe verm." – Aufteilung der Annotationsgruppen

Eine genaue Betrachtung der produzierten MÜ zusammen mit der Gruppenaufteilung zeigt, dass (1) die meisten Sätze der Gruppe RR identisch übersetzt wurden. (2) Vergleicht man die Gruppen FR vs. RF, so kann man argumentieren, dass diese Regel überwiegend hilfreich war. Insbesondere bei den Präfix-Verb-getrennten Fällen konnte die Regel dazu beitragen, MÜ-Fehler zu beheben; so bestand die FR-Gruppe (vor KS) aus 13 Präfix-Verb-getrennten Fällen und 3 Präfix-Verb-ungetrennten Fällen (d. h. Fälle mit untrennbaren Verben oder mit trennbaren Verben, die aber ungetrennt vom Präfix im Satz auftraten).

5.4.9.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Unter allen analysierten KS-Regeln kommt die Regel „Überflüssige Präfixe vermeiden“ auf Platz eins mit den größten Prozentsätzen für die Gruppe RR bei allen Systemen. Unter den MÜ-Systemen waren 88 % der Übersetzungen von dem NMÜ-System Google Translate richtig sowohl mit als auch ohne die Verwendung der Präfixe. Diesem Ergebnis folgen das HMÜ-System Bing mit 71 %; das HMÜ-System Systran mit 67 %; das SMÜ-System SDL mit 63 % und zum Schluss das RBMÜ-System Lucy mit 50 %. Somit war die Gruppe RR die dominanteste Annotationsgruppe unter allen MÜ-Systemen. (siehe Abbildung 5.120)

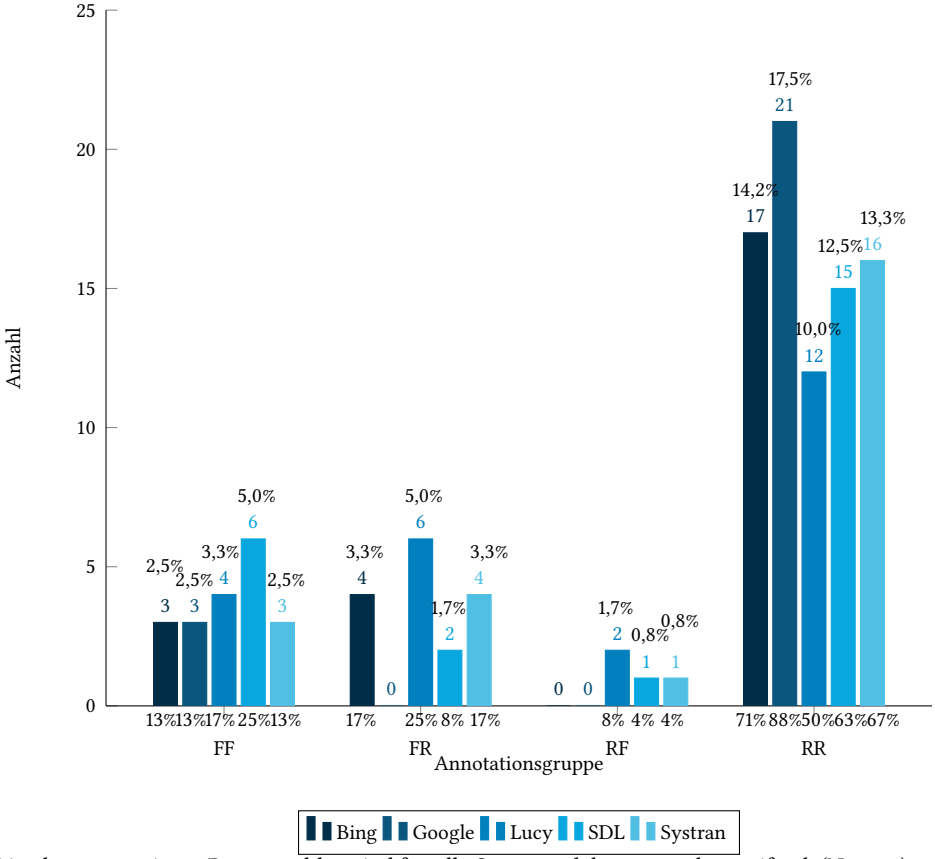
Die zweitgrößte Gruppe war FF: Am höchsten war sie bei SDL mit einem Viertel seiner MÜ repräsentiert (gefolgt von Lucy mit 17 %). Bei den anderen drei Systemen (Bing, Google Translate und Systran) betrug der Anteil dieser Gruppe 13 %. Übersetzungen, die mit überflüssigen Präfixen falsch waren und nach der Regelanwendung korrigiert wurden (Gruppe FR), kamen bei allen Systemen mit Ausnahme des NMÜ-Systems Google Translate vor. Zum Schluss folgt die kleinste Gruppe (RF), die nur für drei Systeme relevant war, nämlich Lucy, SDL und Systran. (siehe Abbildung 5.120) In Tabelle 5.124 vergleichen wir den Output der Gruppen RR und FR dreier Systeme (Systran, Bing und Lucy).

Tabelle 5.124: Beispiel 78

Vor-KS	**Speichern** Sie die angezeigten Werte lokal auf der Festplatte **ab**.
HMÜ Systran	Store the displayed values locally on the hard disk.
HMÜ Bing	Locally, storing the displayed values on the hard disk.
RBMÜ Lucy	Save the displayed values locally on the hard disk.
Nach-KS	**Speichern** Sie die angezeigten Werte lokal auf der Festplatte.
HMÜ Systran	Store the displayed values locally on the hard disk.
HMÜ Bing	Save the displayed values locally on the hard disk.
RBMÜ Lucy	Store the displayed values locally on the hard disk.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Systran und Lucy konnten den Satz mit und ohne überflüssige Präfixe fehlerfrei übersetzen (Gruppe RR). Bei Bing hingegen wurde das Verb mit Präfix (vor



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.120: „Überfl. Präfixe verm.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen

KS) falsch übersetzt und falsch platziert. Erst nach der Regelanwendung konnten die beiden Fehler behoben werden (Gruppe FR). Die Fehlertypen unter dieser Regel werden im folgenden Abschnitt diskutiert.

5.4.9.4 Vergleich der Fehlertypen mit vs. ohne überflüssige Präfixe

Nach Anwendung dieser Regel sank die Fehleranzahl bei dem Fehlertyp LX.4 „Lexik – Zusätzliches Wort eingefügt“ von 9 auf 1 (– 88,8 % / Mv = ,08 / SDv = ,264 / Mn = ,01 / SDn = ,091 / N = 120), siehe Abbildung 5.121. Dieser Unterschied in der Fehleranzahl erwies sich als signifikant (p = ,008 / N = 120). Bei trennbaren Verben steht in manchen Fällen das Präfix am Satzende (vor KS). Hierbei kam es bei einigen Systemen zu einem lexikalischen Fehler, da das Präfix losgelöst vom Satz übersetzt wurde. In solchen Fällen wurde durch das Vermeiden der überflüssigen Präfixe der Fehler behoben (nach KS). Folgende Beispiele beleuchten diesen Fehlertyp: In Tabelle 5.125 übersetzte Systran ‚aus‘ in ‚auswählen‘ zusätzlich am Satzende als ‚out‘.

Tabelle 5.125: Beispiel 79

Vor-KS	**Wählen** Sie die Option "Software von einer bestimmten Liste installieren" **aus**.
HMÜ Systran	Select the option "Install software from a specific list" out.
Nach-KS	**Wählen** Sie die Option "Software von einer bestimmten Liste installieren".
HMÜ Systran	Select the option "Install software from a specific list".

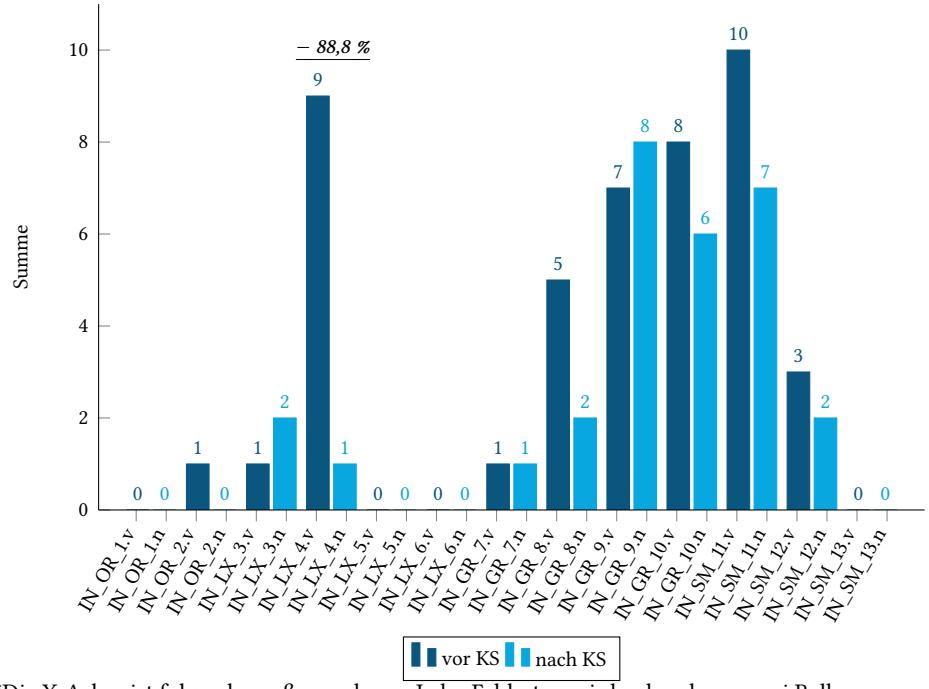
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Ebenfalls übersetzte Lucy – in Tabelle 5.126 – ‚ein‘ in ‚einschicken‘ losgelöst vom Verb als ‚one‘.

Dieser Fehler beeinträchtigte die MÜ-Qualität, wie unter §5.4.9.5 demonstriert. Bei allen anderen Fehlertypen veränderte sich die Fehleranzahl im Vergleich zu nach der Umsetzung der Regel nicht deutlich (Abbildung 5.121).

5.4.9.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt (Abbildung 5.122), dass die Fehleranzahl der einzelnen Fehlertypen auf



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler *vor KS* und der zweite die Summe der Fehler *nach KS*, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
****Signifikante Differenz vor vs. nach KS***
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.121: „Überfl. Präfixe verm.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

Tabelle 5.126: Beispiel 80

Vor-KS	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse **ein**.
RBMÜ Lucy	Please send the appliance in its original packaging to our service address one.
Nach-KS	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse.
RBMÜ Lucy	Please send the appliance in its original packaging to our service address.

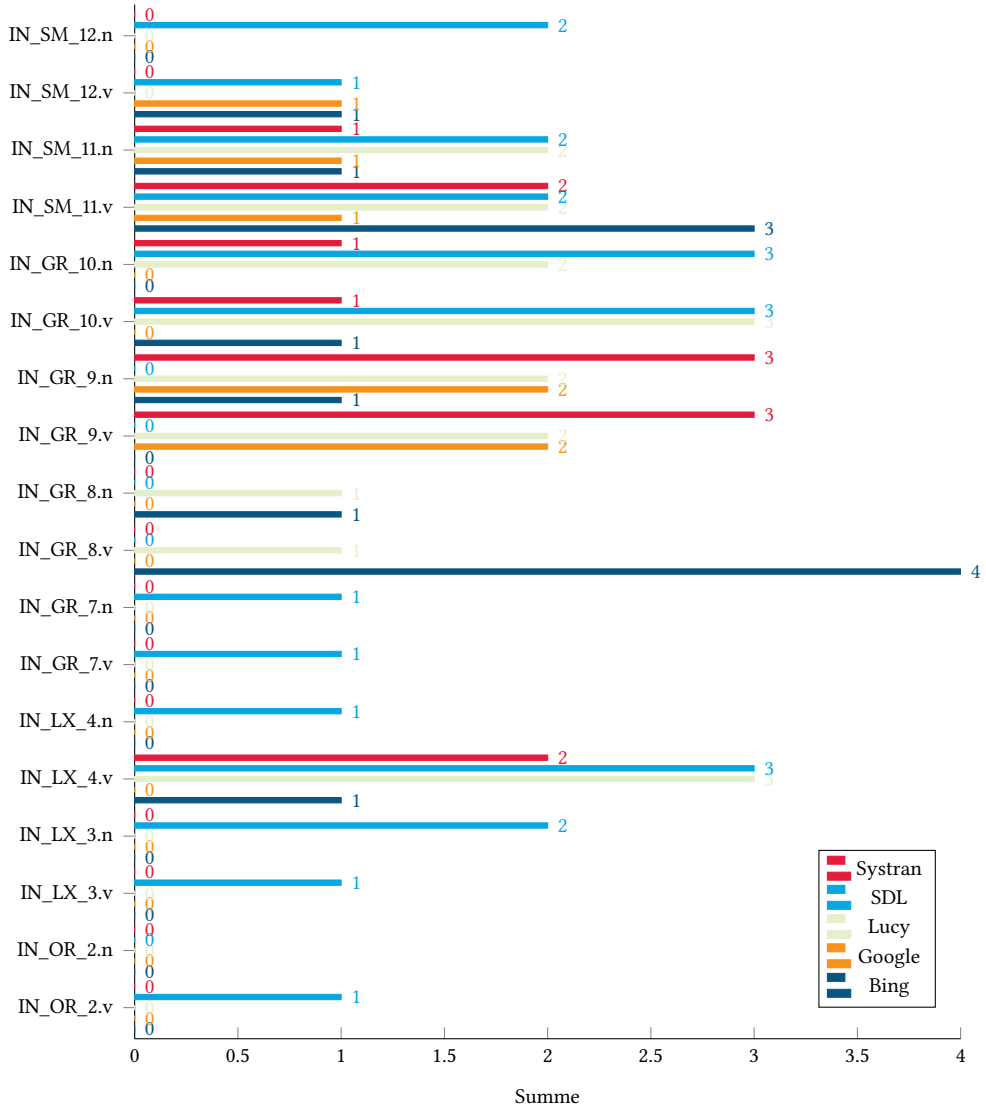
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Systemebene sehr gering war (max. 4). Am meisten beeinflusst durch die Regelanwendung war der Fehlertyp LX.4 „Lexik – Zusätzliches Wort eingefügt“. Dieser Fehlertyp sank bei dem SMÜ-System SDL und wurde bei dem HMÜ-System Bing, dem RBMÜ-System Lucy und dem HMÜ-System Systran vollständig behoben. Jedoch erwies sich die Differenz aufgrund der geringen Fehleranzahl bei keinem der genannten Systeme als signifikant.

5.4.9.5 Vergleich der MÜ-Qualität mit vs. ohne überflüssige Präfixe sowie die Korrelation zwischen den Fehlertypen und der Qualität

Bei der Regel „Überflüssige Präfixe vermeiden“ stiegen zwar die Stil- und Inhaltsqualität,[56] dennoch fiel dieser Anstieg nur sehr niedrig aus. Die meisten Übersetzungen (68 %) bei dieser Regel waren sowohl vor als auch nach der Anwendung der Regel richtig (Gruppe RR), siehe Abbildung 5.119. Daher ist eine statistisch signifikante Qualitätsveränderung schwer vorstellbar. Die Stilqualität stieg um 1,9 % (Mv = 4,25 / SDv = ,554 / Mn = 4,33 / SDn = ,414 / N = 92). Die Inhaltsqualität stieg um 2,5 % (Mv = 4,46 / SDv = ,768 / Mn = 4,57 / SDn = ,701 / N = 92). (siehe Abbildung 5.123) Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag im Fall der Stilqualität bei ,086 (SD = ,435) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,005 und einem Maximum von ,176 sowie im Fall der Inhaltsqualität bei ,111 (SD = ,496) mit einem 95%-Konfidenzintervall zwischen einem Minimum von ,009 und einem

[56] Definitionen der Qualität unter §4.5.5.1.



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 1, 5, 6 und 13 bei gar keinem MÜ-System vor.

Abbildung 5.122: „Überfl. Präfixe verm.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

Maximum von ,214 (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.124. Die Differenzen (nach KS – vor KS) in der Stil- und Inhaltsqualität waren nicht signifikant (p = ,172 bzw. p = ,059).

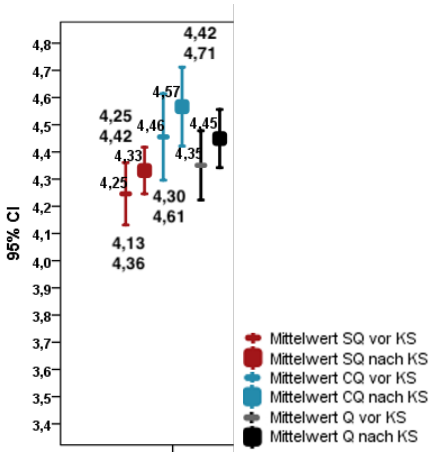


Abbildung 5.123: „Überfl. Präfixe verm.“ – Mittelwerte der Qualität vor und nach KS

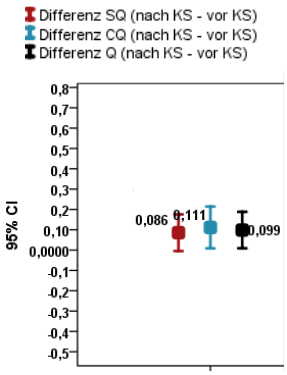


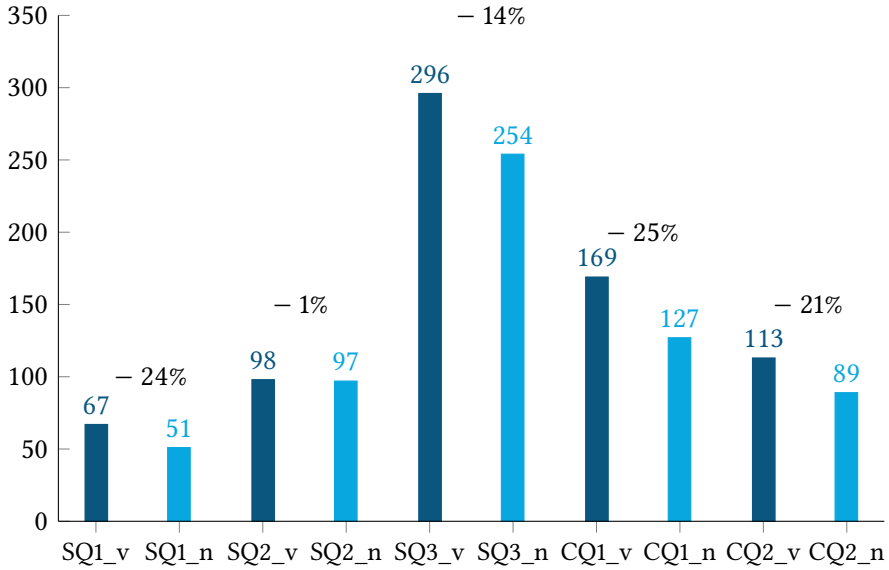
Abbildung 5.124: „Überfl. Präfixe verm.“ - Mittelwert der Qualitätsdifferenzen

Eine genaue Betrachtung der Ergebnisse der Humanevaluation zeigt, dass die Stil- und Inhaltsqualität bei der Behebung der lexikalischen Fehlertyp LX.4 „Zusätzliches Wort eingefügt“ stiegen. Bei der Verwendung von Verben mit überflüssigen Präfixen, wurden die Präfixe zusätzlich übersetzt. Die Behebung dieses Fehlers verbesserte die Übersetzung stilistisch sowie inhaltlich hinsichtlich der Genauigkeit und Verständlichkeit (CQ1 und CQ2).

Dennoch ist der Effekt, wie Tabelle 5.127 zeigt, nur bei dem Präfix bemerkbar. In dem Fall betrug die Qualitätsveränderung auf der Likert-Skala + 0,75 bei der SQ bzw. + 0,63 bei der CQ.

5.4.9.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung bei der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS – Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS – Q. vor KS) zusammenhängt. Mithilfe des Spearman-Tests erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.4 „Zusätzliches Wort eingefügt“ und SM.11 „Verwechslung des Sinns“ einzeln und der Differenz in der Stilqualität; sowie ein



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder.
CQ2: Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.125: „Überfl. Präfixe verm.“ – Vergleich der Qualitätskriterien

Tabelle 5.127: Beispiel 81

Vor-KS	**Wählt** man einen bestimmten Zeichensatz als Standardwert **aus**, wird dieser Zeichensatz in allen Stationen verwendet.
SMÜ SDL	If you select a certain character set as a default value from, this character set will be used in all stations.
Nach-KS	**Wählt** man einen bestimmten Zeichensatz als Standardwert, wird dieser Zeichensatz in allen Stationen verwendet.
SMÜ SDL	If you select a certain character set as a default value, this character set will be used in all stations.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

signifikanter schwacher negativer Zusammenhang zwischen der Differenz im Fehlertyp GR.8 „Falsches Verb (Zeitform, Komposition, Person)“ und der Differenz in der Stilqualität. (siehe Tabelle 5.128)

Tabelle 5.128: „Überfl. Präfixe verm.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	92	< ,001	− ,431
Diff. der Anzahl der **GR.8 „Falsches Verb“**	92	,045	− ,210
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	92	,002	− ,323
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	92	< ,001	− ,434
Diff. der Anzahl der **GR.8 „Falsches Verb“**	92	,005	− ,288
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	92	,001	− ,337
Differenz allg. Q (nach KS − vor KS)			
Diff. der Anzahl der **LX.4 „Zusätzliches Wort eingefügt“**	92	< ,001	− ,452
Diff. der Anzahl der **GR.8 „Falsches Verb“**	92	,008	− ,274
Diff. der Anzahl der **SM.11 „Verwechslung des Sinns“**	92	< ,001	− ,388

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Bezüglich der Inhaltsqualität erwies sich ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in den Fehlertypen LX.4 und SM.11 einzeln und der Differenz in der Inhaltsqualität; sowie ein signifikanter schwacher negativer Zusammenhang zwischen der Differenz im Fehlertyp GR.8 und der

Differenz in der Inhaltsqualität. (siehe Tabelle 5.128) Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht erwiesen werden.

Diese signifikanten negativen Korrelationen deuten darauf hin, dass mit dem Rückgang der Fehleranzahl der genannten Fehlertypen, die Qualität, wie es bei Tabelle 5.127 zu beobachten war, zunahm.

5.4.9.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Die großen Intervalle in Abbildung 5.126 zeigen, dass die Stil- und Inhaltsqualität in den meisten Fällen sehr vergleichbar waren. Durchschnittlich stieg die Qualität bei allen Systemen mit Ausnahme des NMÜ-Systems Google Translate.

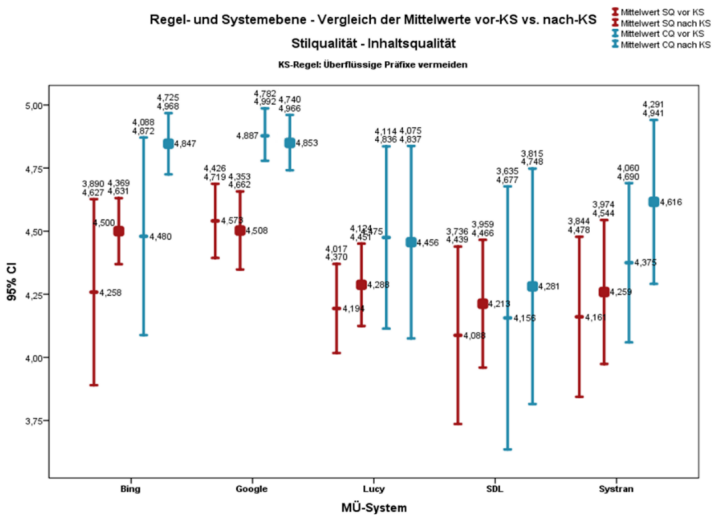


Abbildung 5.126: „Überfl. Präfixe verm." – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

Die geringe Abnahme bei Google Translate von – 1,5 % bei der Stilqualität bzw. – ,7 % bei der Inhaltsqualität entstand durch kleine Unterschiede in Fällen wie Tabelle 5.129.

Obwohl ‚purchase' und ‚buy' Synonyme sind, wurde die MÜ vor KS ‚purchase' nicht nur stilistisch sondern auch inhaltlich höher bewertet (nach KS: – ,63 bei der SQ bzw. – ,25 bei der CQ auf der Likert-Skala). Außerdem wurden weitere Sätze, die vor und nach der Regelanwendung *identisch* übersetzt wurden, durchschnittlich *unterschiedlich* bewertet. Angesichts der allgemein hohen Intrarater-

Tabelle 5.129: Beispiel 82

Vor-KS	**Kaufen Sie die Geräte in einem anderen Land ein, werden Garantieleistungen nur in diesem Land erbracht.**
GNMÜ	If you purchase the devices in another country, warranty services will only be provided in this country.
Nach-KS	**Kaufen Sie die Geräte in einem anderen Land, werden Garantieleistungen nur in diesem Land erbracht.**
GNMÜ	If you buy the devices in another country, warranty services will only be provided in this country.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Agreements (siehe §5.2.4) werfen diese Bewertungsunterschiede eine Forschungsfrage für zukünftige Arbeiten auf. Insgesamt war die Veränderung sowohl in der Stilqualität als auch in der Inhaltsqualität bei keinem der MÜ-Systeme signifikant (Tabelle 5.130).

Tabelle 5.130: „Überfl. Präfixe verm." – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS − vor KS)			Differenz CQ (nach KS − vor KS)			Differenz allg. Q (nach KS − vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	15	,284	− 1,071	15	,074	− 1,785	15	,139	− 1,481
Google	23	,257	− 1,135	23	,475	− ,714	23	,289	− 1,061
Lucy	20	,430	− ,789	20	,959	− ,052	20	,795	− ,259
SDL	20	,154	− 1,425	20	,165	− 1,389	20	,169	− 1,376
Systran	14	,472	− ,719	14	,114	− 1,581	14	,271	− 1,100

p: Signifikanz z: Teststatistik nicht signifikant ($p \geq 0{,}05$)

5.4.9.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwies sich bei dem HMÜ-System Bing nur ein signifikanter starker negativer Zusammenhang zwischen dem grammatischen Fehlertyp GR.8 und der Inhaltsqualität. Bei dem RBMÜ-System Lucy erwies sich eine signifikante starke negative Korrelation zwischen den Fehlertypen GR.8 und SM.11 einzeln und der Stilqualität; sowie eine signifikante starke negative Korrelation zwischen dem Fehlertyp GR.8 und der Inhaltsqualität. Bei dem SMÜ-System SDL erwies sich nur eine signifikante mittlere negative Korrelation zwischen dem Fehlertyp LX.4 und der Stilqualität. Bei dem HMÜ-System Systran erwies sich nur ein signifikanter starker negativer Zusammenhang zwischen dem lexikalischen Fehlertyp LX.4 und der Inhaltsqualität. (siehe Tabelle 5.131)

Tabelle 5.132 zeigt den Effekt der Korrektur des Fehlertyps SM.11 „Verwechslung des Sinns".

In diesem Beispiel wurde der semantische Fehler SM.11 „Verwechslung des Sinns" (in der Übersetzung von ‚absenden' als ‚submit'), nachdem das überflüssige Präfix vermieden wurde, eliminiert. Daraufhin stiegen die Stil- und Inhaltsqualität um 0,13 bzw. 0,63 auf der Likert-Skala (nach KS) an.

5.4.9.6 Vergleich der MÜ-Qualität mit vs. ohne überflüssige Präfixe auf Annotationsgruppenebene

Mit Ausnahme der Gruppe FR war die Differenz in der Stil- und Inhaltsqualität[57] bei allen anderen Annotationsgruppen gering (Abbildung 5.127).

In der Gruppe FF blieb die MÜ bei 16 % der Fälle (siehe §5.4.9.3) aus unterschiedlichen Gründen vor und nach der Regelanwendung falsch. Es gab keinen bestimmten schwerwiegenden Fehler, der bei einem der beiden Szenarien auftrat. Entsprechend war die Qualität vor und nach der Regelanwendung vergleichbar.

In der Gruppe FR stiegen erwartungsgemäß die Stil- und Inhaltsqualität signifikant (Tabelle 5.133): bei der Stilqualität (z (N = 16) = − 3,133 / p = ,002) bzw. bei der Inhaltsqualität (z (N = 16) = − 3,462 / p = ,001).

Oft wurden die getrennt aufgetretenen Präfixe (vor KS) in dieser Gruppe fehlerhaft zusätzlich übersetzt. Die MÜ-Systeme konnten die trennbaren Verben als solche nicht erkennen. Nachdem auf die Präfixe verzichtet wurde (nach KS), wurde die MÜ korrigiert und entsprechend stieg die Qualität. In Tabelle 5.134 war der

[57] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.131: „Überfl. Präfixe verm.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Bing			Lucy			SDL			Systran		
	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ
Differenz SQ (nach KS – vor KS)												
Diff. der Anzahl **LX.4 „Zusätzl. Wort eingef.“**							20	,039	– ,465	14	,120	– ,435
Diff. der Anzahl **GR.8 „Falsches Verb“**	15	,059	– ,498	20	,003	– ,626						
Diff. der Anzahl **SM.11 „Verwechs. d. Sinns“**				20	,022	– ,510						
Differenz CQ (nach KS – vor KS)												
Diff. der Anzahl **LX.4 „Zusätzl. Wort eingef.“**							20	,089	– ,390	14	,020	– ,612
Diff. der Anzahl **GR.8 „Falsches Verb“**	15	,008	– ,658	20	,005	– ,598						
Diff. der Anzahl **SM.11 „Verwechs. d. Sinns“**				20	,227	– ,283						
Differenz Q (nach KS – vor KS)												
Diff. der Anzahl **LX.4 „Zusätzl. Wort eingef.“**							20	,055	– ,435	14	,028	– ,584
Diff. der Anzahl **GR.8 „Falsches Verb“**	15	,012	– ,629	20	,003	– ,623						
Diff. der Anzahl **SM.11 „Verwechs. d. Sinns“**				20	,053	,439						

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (ρ ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Tabelle 5.132: Beispiel 83

Vor-KS	Sorgen Sie dafür, dass die Quelle ein einwandfreies Signal **absendet**.
RBMÜ Lucy	Ensure that the source submits a correct signal.
Nach-KS	Sorgen Sie dafür, dass die Quelle ein einwandfreies Signal **sendet**.
RBMÜ Lucy	Ensure that the source sends a correct signal.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

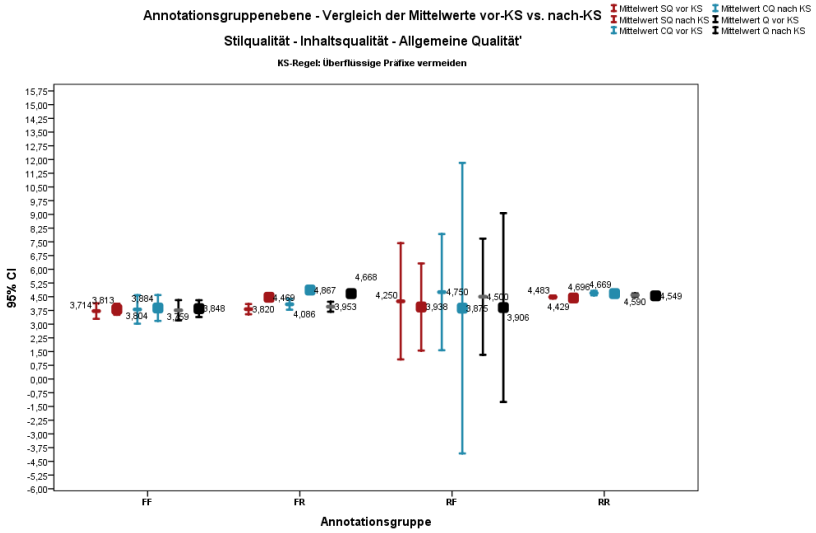


Abbildung 5.127: „Überfl. Präfixe verm.“ – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

Tabelle 5.133: „Überfl. Präfixe verm.“ – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	11	,109	− 1,602
Differenz CQ (nach KS − vor KS)	11	,765	− ,299
Differenz allg. Q (nach KS − vor KS)	11	,439	− ,775
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	16	,002	− 3,133
Differenz CQ (nach KS − vor KS)	16	,001	− 3,462
Differenz allg. Q (nach KS − vor KS)	16	,001	− 3,410
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	2	,180	− 1,342
Differenz CQ (nach KS − vor KS)	2	,180	− 1,342
Differenz allg. Q (nach KS − vor KS)	2	,180	− 1,342
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	62	,148	− 1,448
Differenz CQ (nach KS − vor KS)	62	,828	− ,217
Differenz allg. Q (nach KS − vor KS)	62	,122	− 1,548

Anstieg relativ hoch und betrug bei der Stilqualität 1,00 bzw. bei der Inhaltsqualität 0,88 Punkte auf der Likert-Skala.

Tabelle 5.134: Beispiel 84

Vor-KS	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse **ein**.
RBMÜ Lucy	***Please*** send the appliance in its original packaging to our service address one.
Nach-KS	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse.
RBMÜ Lucy	***Please*** send the appliance in its original packaging to our service address.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

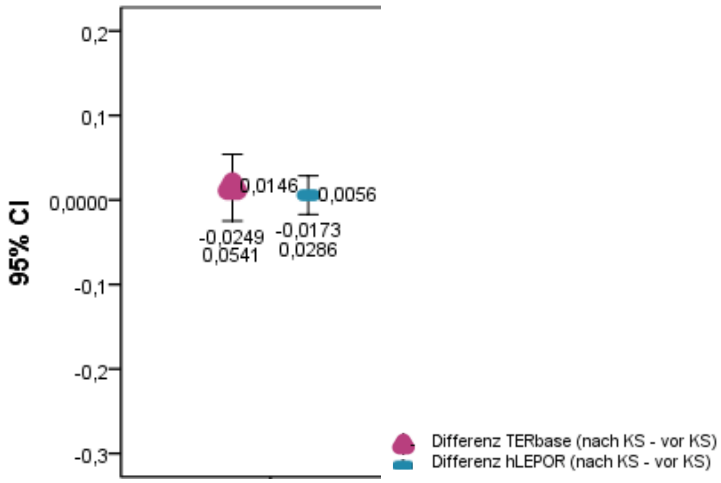
Die Gruppe RF war, wie die Aufteilung der Annotationsgruppen (§5.4.9.3) zeigte, sehr selten vertreten. Nur in 4 Fällen bzw. 3 % der analysierten Sätze und davon wurde nur die Hälfte in der Humanevaluation bewertet, was das relativ große Intervall in der Abbildung 5.127 erklärt.

In der Gruppe RR (Übersetzung vor und nach KS richtig) bleiben die Inhaltsqualität vor und nach der Anwendung der Regel fast unverändert. Hat das System das Präfix richtig geparst, so wurde der Satz fehlerfrei übersetzt und in den meisten Fällen war die MÜ vor und nach der Regelanwendung identisch oder es wurden Synonyme für die Verbübersetzung in den beiden Szenarien verwendet (siehe Tabelle 5.129).

5.4.9.7 Vergleich der AEM-Scores mit vs. ohne überflüssige Präfixe sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores mit vs. ohne die überflüssigen Präfixe zeigte sowohl mit TERbase als auch mit hLEPOR nur eine sehr geringe, nicht signifikante Verbesserung der AEM-Scores, nachdem die Sätze ohne die Präfixe formuliert wurden (nach KS), siehe Abbildung 5.128.

Der Mittelwert der Differenz (nach KS − vor KS) im AEM-Score pro Satz lag für TERbase bei ,015 (SD = ,191) und für die hLEPOR bei ,006 (SD = ,111) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.128. Durch diese minimalen Unterschiede waren die Differenzen (nach KS −



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.128: „Überfl. Präfixe verm." – Mittelwert der Differenz der AEM-Scores

vor KS) in TERbase und hLEPOR nicht signifikant (z (N = 92) = – 1,196 / p = ,232) bzw. (z (N = 92) = – ,735 / p = ,462).

5.4.9.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Nach der Regelanwendung stieg die Qualität leicht (nicht signifikant, §5.4.9.5). Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter mittlerer positiver Zusammenhang zwischen den Differenzen in den AEM-Scores von TERbase und hLEPOR und der Differenz in der allgemeinen Qualität. Bei der Vermeidung von überflüssigen Präfixen verbesserten sich die Scores der beiden AEMs und die Qualität nahm zu. Tabelle 5.135 demonstriert die Korrelationswerte.

Die Korrelationswerte deuten darauf hin, dass die Scores der beiden AEMs sich relativ synchron in die gleiche Richtung wie die allgemeine Qualität bewegten.

5.4.9.8 Analyse der achten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der achten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick

Tabelle 5.135: „Überfl. Präfixe verm." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	92	< ,001	,440	mittlerer Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	92	< ,001	,483	mittlerer Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (−) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[58]

Regel 8: Überflüssige Präfixe vermeiden

Erster Analysefaktor: Vergleich der Fehleranzahl mit vs. ohne überflüssige Präfixe

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

[58] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Resultat

Auf Regelebene:
H0 wurde abgelehnt und somit H1 bestätigt.
Die Fehleranzahl sank signifikant, nachdem die überflüssigen Präfixe vermieden wurden. **Anz.F. (–)**

Auf Regel- und MÜ-Systemebene:
Nur bei Bing sank die Fehleranzahl signifikant nach der Regelanwendung. **Bi (–)**

Bei allen anderen Systemen sank die Fehleranzahl leicht. **Go (–) Lu (–) SD (–) Sy (–)**

Zweiter Analysefaktor

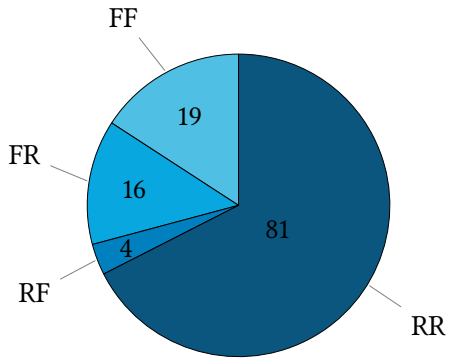


Abbildung 5.129: Aufteilung der Annotationsgruppen auf Regelebene

---

Dritter Analysefaktor: Vergleich der Fehlertypen mit vs. ohne überflüssige Präfixe

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

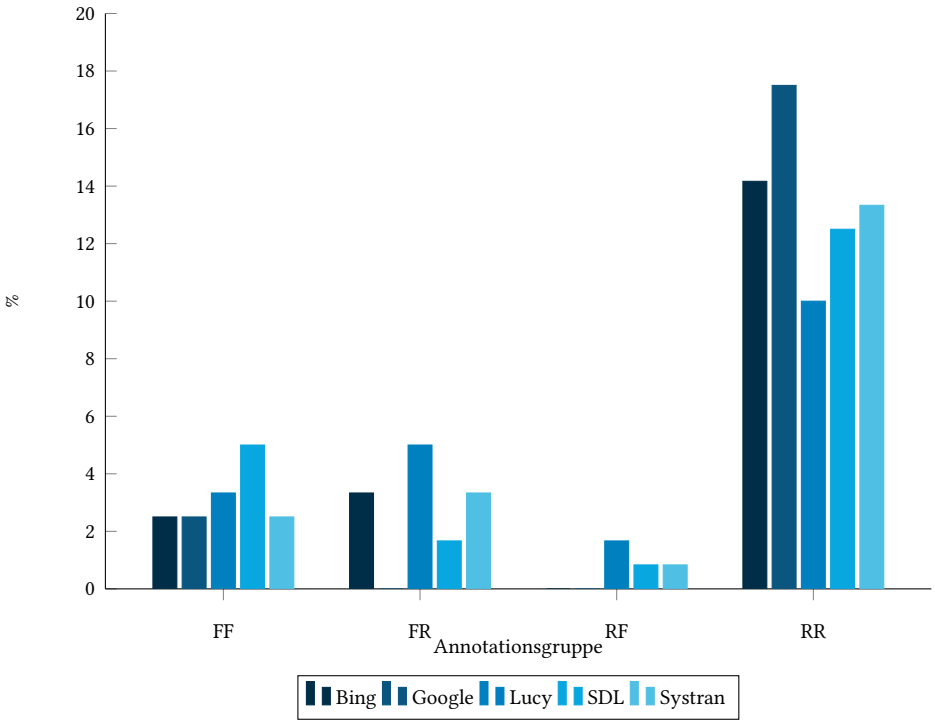


Abbildung 5.130: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Resultat

Auf Regelebene:
H1 wurde nur für einen Fehlertyp bestätigt.
Die Fehleranzahl von LX.4 „Zusätzliches Wort eingefügt“ sank nach der Vermeidung der überflüssigen Präfixe signifikant.
Auf Regel- und MÜ-Systemebene:
Auf Systemebene war die Fehleranzahl sowohl vor als auch nach der Regelanwendung sehr niedrig; es gab keinen bestimmten Fehlertyp, der nach der Regelanwendung statistisch signifikant beeinflusst wurde.

LX.4 (–)

Vierter Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne überflüssige Präfixe

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. **SQ (+)** **CQ (+)**
Die Stil- und Inhaltsqualität stiegen leicht nach der Regelanwendung.
Auf Regel- und MÜ-Systemebene:
Die Differenzen in der Stil- und Inhaltsqualität bei allen Systemen waren (sehr) klein und entsprechend nicht signifikant.

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz in der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS − vor KS) und der Differenz in der Stil- bzw. Inhaltsqualität (Qualität nach KS − vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz in der Fehleranzahl eines bestimmten Fehlertyps und der Differenz in der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz in der Fehleranzahl eines bestimmten Fehlertyps und der Differenz in der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:
H1 wurde für zwei Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und SM.11 „Verwechslung des Sinns“ einzeln und den Differenzen in der Stil- und Inhaltsqualität.

***neg* LX.4 <> SQ**
***neg* SM.11 <> SQ**

***neg* LX.4 <> CQ**
***neg* SM.11 <> CQ**

Auf Regel- und MÜ-Systemebene:
Bei Bing bestand ein signifikanter *starker* negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des GR.8 „Falsches Verb“ und der Differenz in der Inhaltsqualität.

Bi
***neg* GR.8 <<>> CQ**

Bei Lucy bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des GR.8 „Falsches Verb“ und SM.11 „Verwechslung des Sinns“ und der Differenz der Stilqualität sowie ein signifikanter starker negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des GR.8 „Falsches Verb“ und der Differenz in der Inhaltsqualität.

Lu
***neg* GR.8 <<>> SQ**
***neg* SM.11 <<>> SQ**
***neg* GR.8 <<>> CQ**

Bei SDL bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und der Differenz in der Stilqualität.

SD
***neg* LX.4 <> SQ**

Bei Systran bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des LX.4 „Zusätzliches Wort eingefügt“ und der Differenz in der Inhaltsqualität.

Sy
***neg* LX.4 <<>> CQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität mit vs. ohne überflüssige Präfixe auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde nur für die Gruppe FR bestätigt: Bei der Annotationsgruppe FR stiegen die Stil- und Inhaltsqualität signifikant, nachdem das überflüssige Präfix vermieden wurde.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe RF sanken die Stil- und Inhaltsqualität leicht.	**SQ (–)** **CQ (–)**
Bei der Annotationsgruppe FF stiegen die Stil- und Inhaltsqualität minimal.	**SQ (+)** **CQ (+)**
Bei der Annotationsgruppe RR sanken die Stil- und Inhaltsqualität leicht.	**SQ (–)** **CQ (–)**

Siebter Analysefaktor: Vergleich der AEM-Scores mit vs. ohne überflüssige Präfixe

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. Die AEM-Scores von TERbase und hLEPOR verbesserten sich nur leicht, nachdem das überflüssige Präfix vermieden wurde.	**TERbase (+)** **hLEPOR (+)**

Achter Analysefaktor: Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz in den AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz in der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz in den AEM-Scores und der Differenz in der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz in den AEM-Scores und der Differenz in der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen den Differenzen der Scores der beiden AEMs (TERbase und hLEPOR) und der Differenz in der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.10 NEUNTE REGEL: Keine Wortteile weglassen

5.4.10.1 Überblick

Im Folgenden wird die KS-Regel „Keine Wortteile weglassen" kurz beschrieben.[59] Zudem wird zusammenfassend und anhand eines Beispiels demonstriert, wie die Regel bei der Analyse angewendet wurde. Anschließend wird die Aufteilung der Testsätze im Datensatz dargestellt:

Beschreibung der KS-Regel: Keine Wortteile weglassen (tekom-Regel-Nr. S 204)

> Nach dieser Regel sollen die Wörter vollständig geschrieben werden (tekom 2013: 68).
>
> Begründung: Ziel der Regel ist die Unterstützung des Verständnisses. Insbesondere bei der Übersetzung ist das Weglassen von Wortteilen ungeeignet. (ebd.)

Umsetzungsmuster:

Vor KS: Der Satz beinhaltet Wortteile.
Nach KS: Die fehlenden Wortteile werden vervollständigt.

[59] Die für diese Regel relevanten Kontraste im Sprachenpaar DE-EN sind unter §4.5.2.3 erörtert.

KS-Stelle

Vor KS: Wörter mit den weggelassenen Teilen
Nach KS: vollständige Wörter

Beispiele

Die Ist- und Sollwerte des zweiten Regelkreises werden nach der Konfiguration angezeigt.

Der Istwert und der Sollwert des zweiten Regelkreises werden nach der Konfiguration angezeigt.

Aufteilung der Testsätze: Der Datensatz besteht aus 24 verschiedenen Begriffen mit unterschiedlichen Wortteilen. Die Geläufigkeit der untersuchten Begriffe wurde anhand der Anzahl der Treffer bei einer Google-Suche gemessen. Die Aufteilung der untersuchten Begriffe war wie folgt:

1 Begriff mit > 1.000.000 Treffern;

3 Begriffe mit > 100.000 und < 1.000.000 Treffern;

7 Begriffe mit > 1.000 und < 100.000 Treffern;

13 Begriffe mit < 1.000 bis 0 Treffer.

Im Weiteren werden die Ergebnisse der einzelnen Analysefaktoren präsentiert.

5.4.10.2 Vergleich der Fehleranzahl vor vs. nach dem Weglassen von Wortteilen

Die Fehleranzahl sank minimal um 6,6 % von 76 Fehlern bei dem Weglassen von Wortteilen (M = ,63 / SD = ,788 / N = 120) auf 71 Fehler bei der Verwendung von vollständigen Wörtern (M = ,59 / SD = ,815 / N = 120), siehe Abbildung 5.131. Der Mittelwert der Differenz (nach KS − vor KS) in der Fehleranzahl pro Satz lag somit bei − ,04 (SD = ,864) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,21 (SD = ,700) und einem Maximum von ,12 (SD = 1,014) (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.132. Die Differenz (nach KS − vor KS) in der Fehleranzahl war entsprechend nicht signifikant (z (N = 120) = − ,445 / p = ,656).

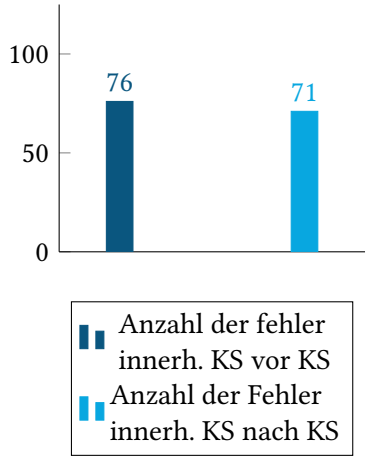


Abbildung 5.131: „Keine Wortteile wegla." – Fehlersumme vor vs. nach KS

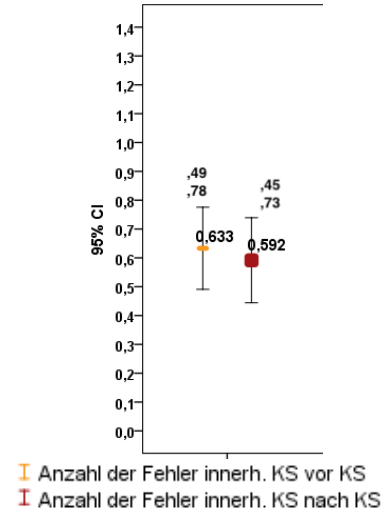


Abbildung 5.132: „Keine Wortteile wegla." – Mittelwert der Fehleranzahl pro Satz vor vs. nach KS

Es wurde bei dieser Regel angenommen, dass der Geläufigkeitsgrad der abgekürzten Begriffe eine Rolle bei der Korrektheit von deren Übersetzung spielt.[60] Die Grundidee dieser Annahme ist, dass gebräuchliche abgekürzte Begriffe in ihrer abgekürzten Form in den MÜ-Systemen lexikalisch hinterlegt sind oder in den Trainingsdaten vorkommen). Dementsprechend wäre in diesem Fall die Wahrscheinlichkeit einer korrekten Übersetzung hoch und die Anwendung der Regel nicht erforderlich. Bei ungebräuchlichen abgekürzten Begriffen hingegen wurde erwartet, dass die Verwendung der vollständigen Wörter (d. h. die Anwendung der Regel) eine korrekte MÜ fördert. Der Geläufigkeitsgrad der Begriffe wurde im Sinne der Anzahl der Treffer bei einer Google-Suche ermittelt (Tabelle 5.136).

Tabelle 5.136 zeigt die Aufteilung des Datensatzes (24 Sätze x 5 MÜ) nach dem Geläufigkeitsgrad der untersuchten Begriffe sowie die Differenz in der Fehleranzahl bei jeder Gruppe. Alle untersuchten Begriffe hatten im Suchvorgang deutlich mehr Treffer in ihrer abgekürzten Form als in ihrer vollständigen. Bei einer hohen Geläufigkeit (4 Fälle) gab es keine Differenz in der Fehleranzahl. Die

[60] Da die Studie generische Black-Box-Systeme untersuchte, bei denen keine Terminologieintegration erfolgen konnte, wurden firmen- und produktspezifische Termini durch geläufige Begriffe ersetzt. Für die Auswahlkriterien der untersuchten Systeme siehe §4.5.1. Für den genauen Umgang mit den spezifischen Termini im Rahmen der Studie siehe Schritt [4] unter §4.5.3.1.

Tabelle 5.136: Daten und Fehleranzahlveränderung der untersuchten Fälle

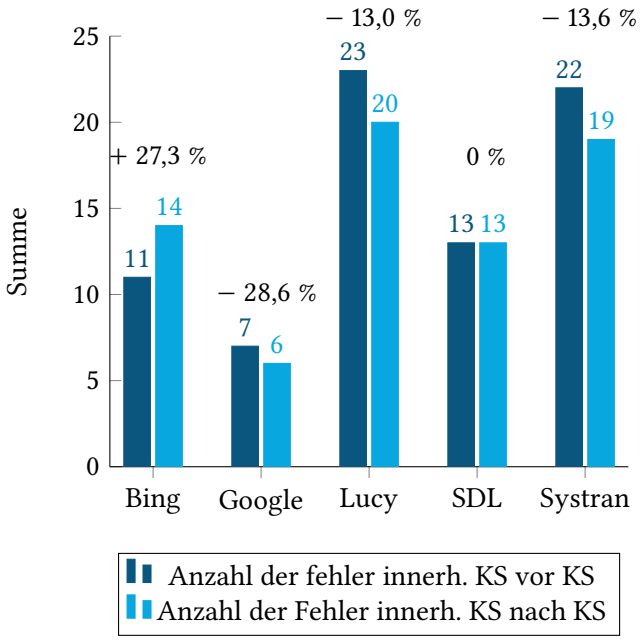
	Abgekürzte Begriffe nach ihren Geläufigkeiten			
	hoch		**moderat**	**niedrig**
Anz. Treffer Google-Suche	> 1.000.000	> 100.000 und < 1.000.000	> 1.000 und < 100.000	< 1.000 bis 0
Anzahl der Fälle	1 x 5 MÜ	3 x 5 MÜ	7 x 5 MÜ	13 x 5 MÜ
Durchschnittliche Diff. F.Anz. (nach KS − vor KS)	0	0	0,29	− ,69

MÜ-Systeme konnten die Begriffe sowohl in ihrer abgekürzten als auch in ihrer vollständigen Form fehlerfrei übersetzen. Wie die Qualitätswerte in den beiden fehlerfreien Szenarien ausfielen, wird unter §5.4.10.5 erläutert. Bei einer moderaten Geläufigkeit (7 Fälle) stieg die Fehleranzahl bei der Verwendung der vollständigen Form der Wörter (nach KS). Ein Beispiel dieser Fälle ist der Begriff ‚Bedienungs- und Pflegehinweise'. Dieser Begriff hat bei der Google-Suche 6.100 Treffer in seiner abgekürzten Form bzw. 9 Treffer in seiner vollständigen Form ergeben. Da diese Gruppe der Begriffe eher in ihrer abgekürzten Form geläufig ist, stieg die Fehleranzahl bei der Verwendung der vollständigen Form (nach KS) (durchschnittlicher Anstieg von 0,29). Bei einer sehr niedrigen Geläufigkeit der Begriffe (13 Fälle) handelt es sich um Begriffe, die in ihrer abgekürzten Form nicht geläufig sind. Hierbei war die Annahme, dass die Verwendung der vollständigen Form die MÜ erleichtern würde (z. B. ‚Soja- und laktosefreie Milch' als ‚Sojamilch und laktosefreie Milch'). Tatsächlich sank die Fehleranzahl bei diesen Fällen nach der Regelanwendung (durchschnittlicher Rückgang von − ,69). Aufgrund der kleinen Anzahl der Fälle jeder Gruppe bedarf jedoch die Annahme über einen Zusammenhang zwischen dem Geläufigkeitsgrad und der Korrektheit der MÜ weiterer Forschung.

5.4.10.2.1 Vergleich der Fehleranzahl auf Regel- und MÜ-Systemebene

Auf Systemebene war die Veränderung in der Fehleranzahl insignifikant. Bei drei Systemen sank die Fehleranzahl (Abbildung 5.133): NMÜ-System Google Translate (Mdiff = − 0,083); HMÜ-System Systran (Mdiff = − 0,125); RBMÜ-System Lucy (Mdiff = − 0,125). Bei dem SMÜ-System SDL blieb die Fehleranzahl unverändert.

Das HMÜ-System Bing war das einzige System, bei dem die Fehleranzahl stieg (Mdiff = 0,125).



Signifikante Differenz vor vs. nach KS

Abbildung 5.133: „Keine Wortteile wegla." – Summe der Fehleranzahl vor vs. nach KS bei den einzelnen MÜ-Systemen

Tabelle 5.137 bietet einen Vergleich zwischen der MÜ von SDL und Google Translate an.

Wie das Beispiel zeigt, konnte SDL die abgekürzte Form des Begriffs ‚Anwärmvorgang' vor der Regelanwendung übersetzen, während die vollständige Form desselben Begriffs unübersetzt blieb. Google Translate hingegen konnte beide Versionen fehlerfrei übersetzen. Inwiefern die Qualität der MÜ in den beiden Fällen akzeptabel ist, wird unter §5.4.10.5 diskutiert.

5.4.10.3 Aufteilung der Annotationsgruppen

Wie Abbildung 5.134 demonstriert, waren knapp 43 % der Übersetzungen sowohl vor als auch nach dem Weglassen der Wortteile fehlerfrei (Gruppe RR). Gleichzeitig war die zweitgrößte Gruppe FF (ca. 28 %) mit MÜ, die sowohl vor als auch nach dem Weglassen der Wortteile Fehler beinhalteten. Die Fehlertypen werden genauer unter §5.4.10.4 analysiert.

Tabelle 5.137: Beispiel 85

Vor-KS	Die **Anwärm- und Entwässerungsvorgänge** sind gemäß der Betriebsanleitung zu beachten.
SMÜ SDL	The warm up and drainage operations have to be observed in accordance with the operating instructions.
GNMÜ	The heating and de-watering procedures have to be observed in accordance with the operating instructions.
Nach-KS	Der **Anwärmvorgang und der Entwässerungsvorgang** sind gemäß der Betriebsanleitung zu beachten.
SMÜ SDL	The Anwarmvorgang and drainage operations have to be observed in accordance with the operating instructions.
GNMÜ	The heating process and the dewatering process have to be observed in accordance with the operating instructions.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Abbildung 5.134: „Keine Wortteile wegla.“ – Aufteilung der Annotationsgruppen

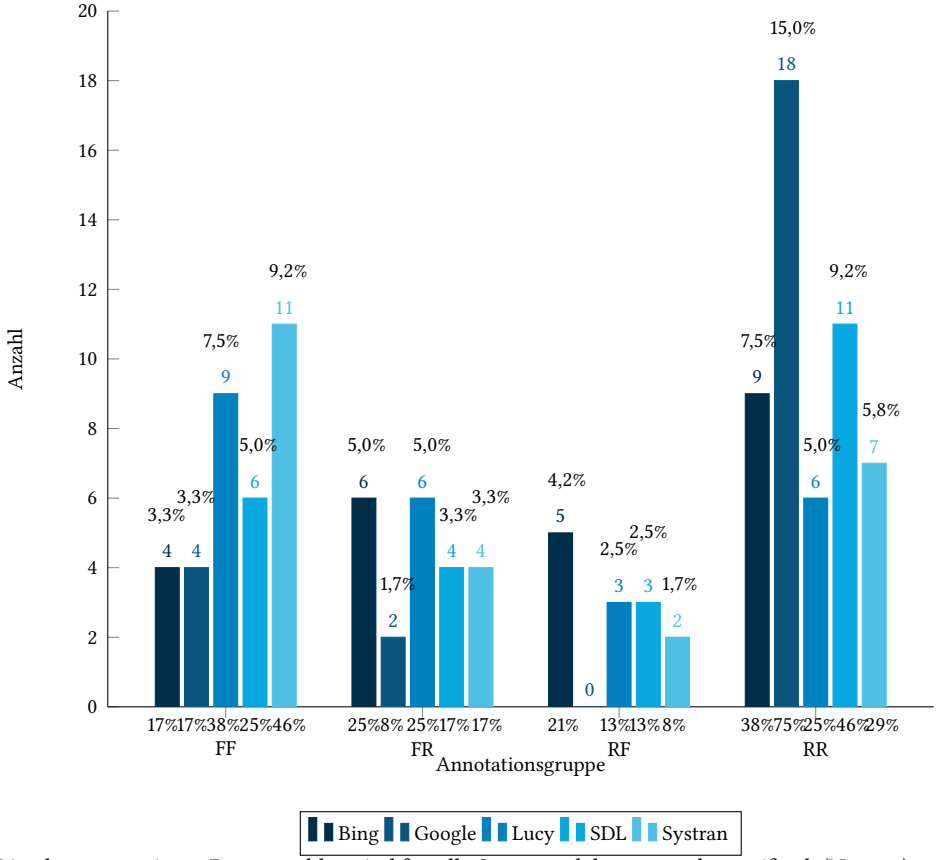
Bei ca. 18 % der Fälle (Gruppe FR) unterstützte die Regel die MÜ-Systeme dabei, die vor KS aufgetretenen Fehler zu eliminieren. Gleichzeitig traten bei knapp 11 % der Fälle Fehler erst nach der Regelanwendung auf (Gruppe RF). Inwiefern das Auftreten bzw. die Eliminierung von Fehlern die MÜ-Qualität beeinflusste, wird unter §5.4.10.5 und §5.4.10.6 diskutiert.

5.4.10.3.1 Vergleich der Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Die Gruppe RR war bei drei Systemen die größte, so war sie bei dem NMÜ-System Google Translate mit 75 %, dem SMÜ-System SDL mit 46 % und dem HMÜ-Systems Bing mit 38 % repräsentiert (Abbildung 5.135). Diese drei Systeme waren in der Lage, die genannten Anteile der Sätze sowohl bei Verwendung der abgekürzten Form als auch der vollständigen Form der untersuchten Begriffe fehlerfrei zu übersetzen. Auf der anderen Seite war die Gruppe FF hoch repräsentiert bei dem HMÜ-System Systran mit 46 % und dem RBMÜ-System Lucy mit 38 %. Gleichzeitig – wie Abbildung 5.135 zeigt – ist das Ergebnis bei allen Systemen mit Ausnahme des NMÜ-Systems Google Translate sehr gemischt. Nur bei Google Translate gab es gar keine Sätze, die vor der Regelanwendung richtig übersetzt wurden und nachher falsch (Gruppe RF).

Zu beobachten ist auch, dass die Gruppe FR größer als RF war. Welche Fehlertypen nach der Regelanwendung auftraten bzw. eliminiert wurden, werden unter §5.4.10.4 dargestellt. Die Gruppe RF war – vor allem bei Bing – relativ hoch vertreten. Betrachten wir Tabelle 5.138, in dem Bing nur nach der Regelanwendung eine falsche MÜ lieferte, während Google Translate in beiden Szenarien fehlerfrei übersetzen konnte.

‚Spannungs- und Druckquellen' ist ein Begriff, der ungeläufig ist. In einer Google-Suche gab es gar keinen Treffer weder für diese abgekürzte Form (vor KS) noch für die vollständige Form ‚Spannungsquellen und Druckquellen' (nach KS). Die Verwendung der vollständigen Form zeigte sich nicht sinnvoll im Fall von Bing. Es trat ein semantischer Fehler auf (‚print' anstelle von ‚pressure' als Übersetzung für ‚Druck'). Dieser Fehler kann in der Praxis durch eine Terminologieintegration vermieden werden. Gleichzeitig konnte Google Translate beide Formen fehlerfrei übersetzen. Wie die Qualitätsanalyse unter §5.4.10.5 mehr verrät, sanken die Qualitätswerte bei den beiden Systemen nach der Regelanwendung, wie es der Fall in Tabelle 5.138 ist.



Die oben angezeigten Prozentzahlen sind für alle Systeme, d. h. systemübergreifend, (N = 120) berechnet.
Die untenstehenden Prozentzahlen sind auf Systemebene (N = 24) berechnet.

Abbildung 5.135: „Keine Wortteile wegla.“ – Aufteilung der Annotationsgruppen bei den einzelnen MÜ-Systemen

Tabelle 5.138: Beispiel 86

Vor-KS	Trennen Sie alle **Spannungs- und Druckquellen** von der Maschine.
HMÜ Bing	Disconnect all voltage and pressure sources from the machine.
GNMÜ	Disconnect all voltage and pressure sources from the machine.
Nach-KS	Trennen Sie alle **Spannungsquellen und Druckquellen** von der Maschine.
HMÜ Bing	Disconnect all power sources and print sources from the machine.
GNMÜ	Disconnect all voltage sources and pressure sources from the machine.

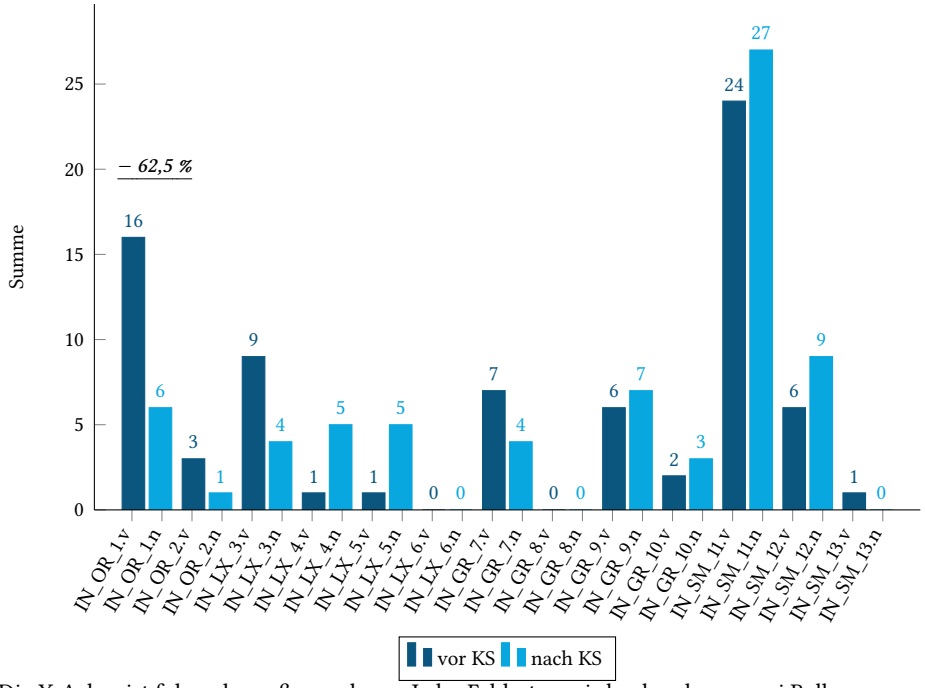
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

5.4.10.4 Vergleich der Fehlertypen vor vs. nach dem Weglassen von Wortteilen

Nach der Regelanwendung sank die Fehleranzahl bei dem Fehlertyp OR.1 „Orthografie – Zeichensetzung“ von 16 auf 6 (– 62,5 % / Mv = ,13 / SDv = ,409 / Mn = ,05 / SDn = ,254 / N = 120), siehe Abbildung 5.136. Der Unterschied in der Fehleranzahl erwies sich als signifikant (p = ,039 / N = 120). Bei allen anderen Fehlertypen veränderte sich die Fehleranzahl nach der Regelanwendung nicht deutlich.

Die Verwendung des Bindestriches (bzw. des Ergänzungsstriches) beim Weglassen von Wortteilen (vor KS) war mit einer falschen Zeichensetzung verbunden. Die Regeln der Bindestrichsetzung im Deutschen und im Englischen unterscheiden sich. Daher war eine korrekte Bindestrichsetzung (vor KS) in manchen Fällen problematisch. Nachdem vollständige Wörter verwendet wurden (nach KS), wurde dieser Fehler in mehreren Übersetzungen behoben. Tabelle 5.139 demonstriert diesen Fall.

Im Englischen wird der Bindestrich in Verbindung mit Adjektiven, nicht mit Substativen, wie es hier der Fall im Deutschen ist, verwendet. Daher trat der Zeichensetzungsfehler (OR.1) in ‚Soya-‘ auf. Zudem trat ein weiterer orthografischer Fehler auf, nämlich der Großschreibungsfehler (OR.2) ebenfalls in dem Wort ‚Soya‘. Nach der Regelanwendung wurden beide Fehler behoben.



*Die X-Achse ist folgendermaßen zu lesen: Jeder Fehlertyp wird anhand von zwei Balken abgebildet. Der erste Balken repräsentiert die Summe der Fehler *vor KS* und der zweite die Summe der Fehler *nach KS*, somit steht z. B. „OR_1.v“ für „OR_1: orthografischer Fehler Nr. 1“ und „v: vor KS“; „OR_1.n“ wäre entsprechend das Pendant zu „OR_1.v“ für das nach-KS-Szenario („n“).
****Signifikante Differenz vor vs. nach KS***
OR.1: Orthografie – Zeichensetzung
OR.2: Orthografie – Großschreibung
LX.3: Lexik – Wort ausgelassen
LX.4: Lexik – Zusätzliches Wort eingefügt
LX.5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
LX.6: Lexik – Konsistenzfehler
GR.7: Grammatik – Falsche Wortart / Wortklasse
GR.8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
GR.9: Grammatik – Kongruenzfehler (Agreement)
GR.10: Grammatik – Falsche Wortstellung
SM.11: Semantik – Verwechslung des Sinns
SM.12: Semantik – Falsche Wahl
SM.13: Semantik – Kollokationsfehler

Abbildung 5.136: „Keine Wortteile wegla.“ – Summe der Fehleranzahl der einzelnen Fehlertypen vor vs. nach KS

Tabelle 5.139: Beispiel 87

Vor-KS	Sogar **Soja- und laktosefreie Milch** lassen sich mit dieser Maschine perfekt aufschäumen.
RBMÜ Lucy	Even Soya- and lactose-free milk can be perfectly frothed with this machine.
Nach-KS	Sogar **Sojamilch und laktosefreie Milch** lassen sich mit dieser Maschine perfekt aufschäumen.
RBMÜ Lucy	Even soya milk and lactose-free milk can be perfectly frothed with this machine.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

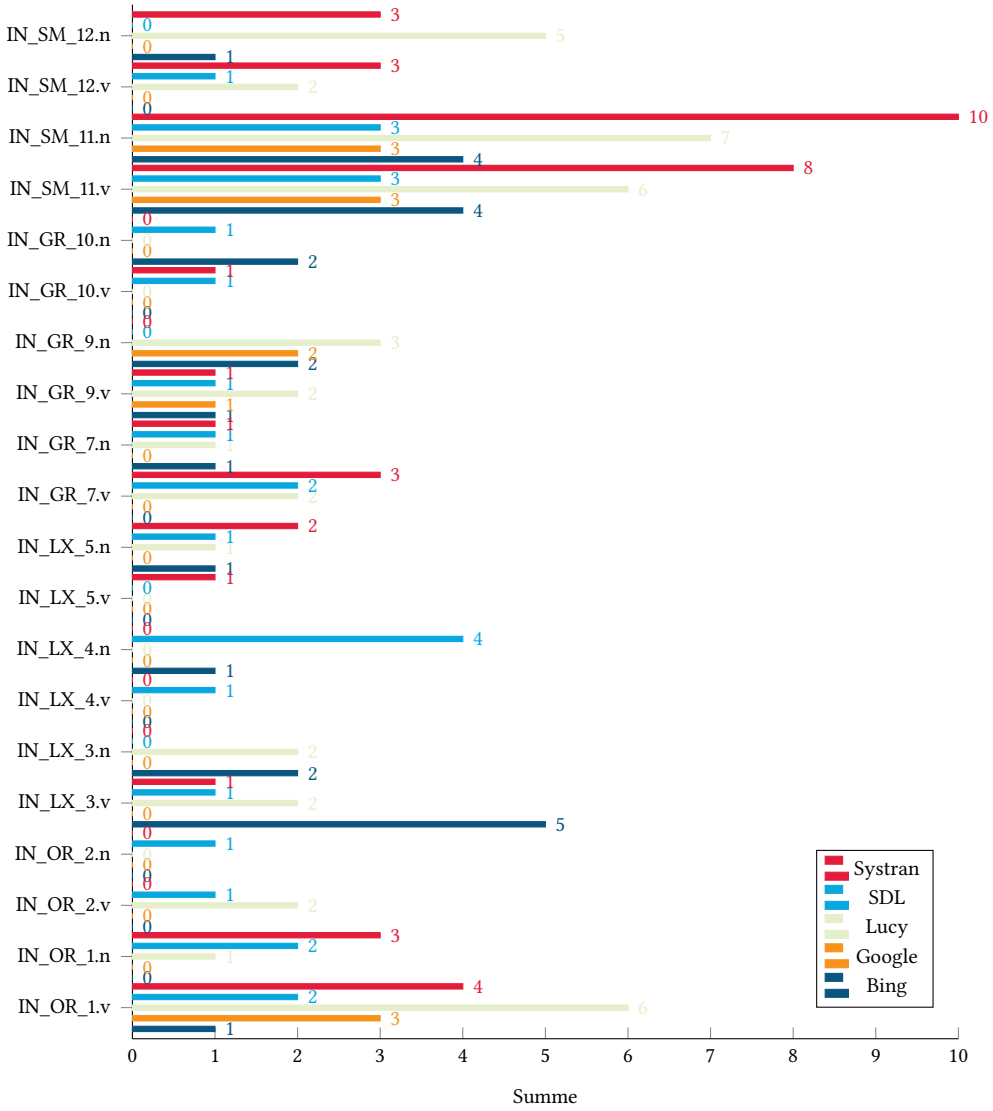
5.4.10.4.1 Vergleich der Fehlertypen auf Regel- und MÜ-Systemebene

Eine genauere Untersuchung der Fehlertypen bei den verschiedenen MÜ-Systemen zeigt (Abbildung 5.137), dass Fehlertyp OR.1 „Zeichensetzung“ bei dem RBMÜ-System Lucy und dem HMÜ-System Systran sank und bei dem HMÜ-System Bing und dem NMÜ-System Google Translate vollständig behoben wurde.

Die Fehleranzahl beim Fehlertyp OR.1 „Zeichensetzung“ war jedoch nicht hoch und erwies sich bei keinem der genannten Systeme als signifikant. Weitere deutliche Veränderungen bei den anderen Fehlertypen waren ebenfalls nicht zu beobachten.

5.4.10.5 Vergleich der MÜ-Qualität vor vs. nach dem Weglassen von Wortteilen sowie die Korrelation zwischen den Fehlertypen und der Qualität

Wie unter §5.4.10.2 dargestellt, war eine moderate Geläufigkeit der Begriffe in ihrer abgekürzten Form mit einer niedrigeren Fehleranzahl (im Vergleich zu der vollständigen Form) verbunden (Tabelle 5.140). Auf der anderen Seite zeigten die Fälle mit einer sehr niedrigen Geläufigkeit in der abgekürzten Form eine Verbesserung im Sinne eines Rückgangs der Fehleranzahl bei der Verwendung der vollständigen Form (nach KS) (Tabelle 5.140). Trotz der kleinen Anzahl der Fälle jeder Gruppe zeigt dieses Ergebnis eine Tendenz, dass je geläufiger ein Begriff in seiner abgekürzten Form war, desto fehlerfreier war auch seine MÜ (d. h. eine Anwendung der Regel war nicht erforderlich). Umgekehrt konnte die Regelanwendung



*Die Balken zeigen die Summe der Fehleranzahl bei jedem Fehlertyp, wobei „v“ für die Summe „vor der Anwendung der KS-Regel“ und „n“ für die Summe „nach der Anwendung der KS-Regel“ steht. Jeder Fehlertyp wird erst für alle Systeme für das Szenario „vor KS“ abgebildet, danach folgt derselbe Fehlertyp wieder für alle Systeme für das Szenario „nach KS“.
**Um die Übersichtlichkeit und Lesbarkeit der Grafik zu erhöhen, wurden in der Grafik die Fehlertypen ausgeblendet, die 0 oder nur einmal bei *allen* MÜ-Systemen vorkamen: In dieser Grafik kamen die Fehlertypen 6 und 8 bei gar keinem MÜ-System vor. Zudem kam der Fehlertyp 13 nur einmal jeweils bei einem MÜ-System vor.

Abbildung 5.137: „Keine Wortteile wegla.“ – Summe der Fehleranzahl der Fehlertypen vor vs. nach KS bei den einzelnen MÜ-Systemen

bei ungeläufigen abgekürzten Begriffen dazu beitragen, die Fehleranzahl zu reduzieren. Eine Reduzierung der Fehleranzahl deutet nicht zwangsläufig auf eine verbesserte Qualität hin. Auf Basis der Ergebnisse der Humanevaluation sanken die Stil- und Inhaltsqualität[61] überall nach der Regelanwendung unabhängig davon, ob die Begriffe in ihrer abgekürzten Form geläufig oder ungeläufig waren (Tabelle 5.140).

Tabelle 5.140: Qualitätsveränderung bei den untersuchten Fällen

	Abgekürzte Begriffe nach ihren Geläufigkeiten							
	hoch				**moderat**		**niedrig**	
Anz. Treffer Google-Suche	> 1.000.000		> 100.000 und < 1.000.000		> 1.000 und < 100.000		< 1.000 bis 0	
Anzahl der Fälle	1 x 5 MÜ		3 x 5 MÜ		7 x 5 MÜ		13 x 5 MÜ	
Durchschnittliche Diff. F.Anz. (nach KS − vor KS)	0		0		0,29		− ,69	
Durchschnittliche Qualitäts-veränderung (nach KS − vor KS)	SQ − ,41	CQ − ,19	SQ − ,25	CQ − ,16	SQ − ,25	CQ − ,16	SQ − ,18	CQ − ,15

Der Einfluss auf die Stilqualität war größer im Vergleich zur Inhaltsqualität (Abbildung 5.139). Die Stilqualität sank um 5,7 % (Mv = 4,23 / SDv = ,653 / Mn = 3,99 / SDn = ,557 / N = 87). Die Inhaltsqualität sank um 3,5 % (Mv = 4,29 / SDv = ,745 / Mn = 4,14 / SDn = ,800 / N = 87). Der Mittelwert der Differenz (nach KS − vor KS) der vergebenen Qualitätspunkte pro Satz lag im Fall der Stilqualität bei − ,246 (SD = ,536) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,360 und einem Maximum von − ,131 und im Fall der Inhaltsqualität bei − ,158 (SD = ,781) mit einem 95%-Konfidenzintervall zwischen einem Minimum von − ,325 und einem Maximum von ,008 (Bootstrapping mit 1000 Stichproben), siehe Abbildung 5.139. Nur die Differenz (nach KS − vor KS) in der Stilqualität erwies sich als hochsignifikant (z (N = 87) = − 4,367 / p < ,001). Bei der Inhaltsqualität war die Differenz insignifikant (z (N = 87) = − 1,764 / p = ,078).

Die Regelanwendung war aufgrund der Wiederholung der Wortteile mit einer stilistischen Inakzeptanz verbunden. An dieser Stelle muss wiederholt erwähnt

[61] Definitionen der Qualität unter §4.5.5.1.

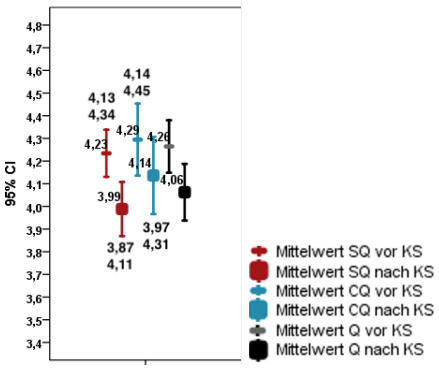


Abbildung 5.138: „Keine Wortteile wegla." – Mittelwerte der Qualität vor und nach KS

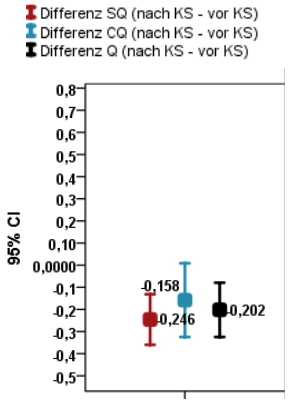


Abbildung 5.139: „Keine Wortteile wegla." – Mittelwert der Qualitätsdifferenzen

werden, dass die Studie mithilfe generischer Black-Box-Systeme und entsprechend ohne Terminologieintegration durchgeführt wurde.[62] In der Praxis haben die Unternehmen die Möglichkeit, bestimmte Begriffe als Termini einzupflegen und in ihre MÜ-Systeme zu integrieren, sodass bei einer MÜ je nach den festgelegten Termini übersetzt wird. Eine Anwendung dieser Regel kann daher bei kritischen Inhalten (wie z. B. bei Warnhinweisen) zwecks Klarheit sinnvoll sein, auch wenn der Text stilistisch nicht ideal ist. Je nach Kontext und Satzintention ist eine Abwägung zwischen Klarheit und stilistischer Akzeptanz erforderlich.

In Tabelle 5.141 sanken sowohl die Stil- als auch die Inhaltsqualität nach der Regelanwendung (– 1,25 bei der Stilqualität bzw. – ,75 Punkte auf der Likert-Skala bei der Inhaltsqualität), obwohl der Satz mit und ohne die Verwendung von vollständigen Wörtern (d. h. vor und nach KS) fehlerfrei übersetzt wurde. Die Bewerter fanden die MÜ mit Wortteilen (vor der Anwendung der KS-Regel) prägnanter und idiomatischer im Vergleich zur vollständigen Form in ‚Eingangskonfiguration und Ausgangskonfiguration' (nach KS).

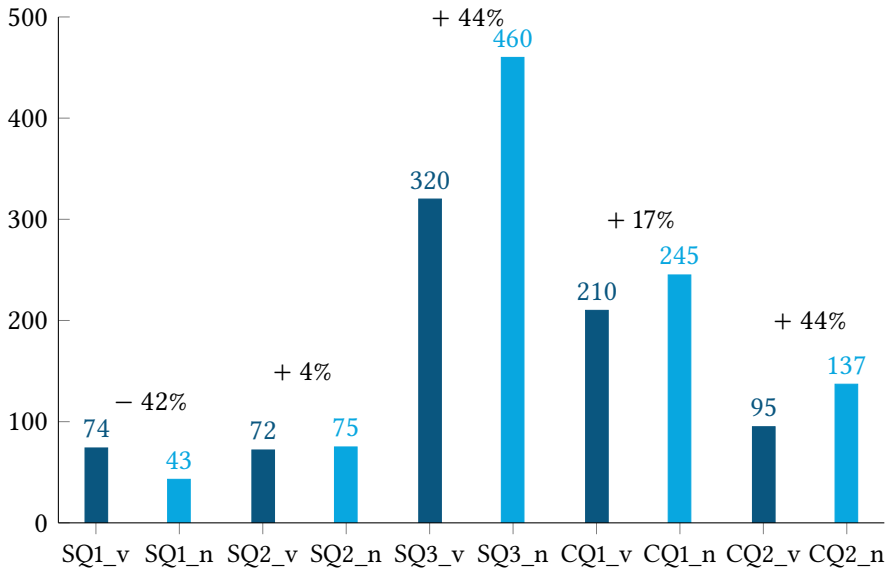
Wie Abbildung 5.140 zeigt, lag der Rückgang in der Stilqualität überwiegend an der mangelnden Idiomatik der Formulierung (SQ3) sowie bei der Inhaltsqualität insbesondere an der beeinträchtigen Verständlichkeit (CQ2). Für eine bessere Vorstellung dieses Rückgangs betrachten wir Tabelle 5.142, bei dem der Satz vor und nach der Formulierung mit Wortteilen falsch übersetzt wurde.

[62] Für die Auswahlkriterien der untersuchten Systeme siehe §4.5.1. Für den genauen Umgang mit den spezifischen Termini im Rahmen der Studie siehe Schritt [4] unter §4.5.3.1.

Tabelle 5.141: Beispiel 88

Vor-KS	Die wichtigsten Parameter der **Ein- und Ausgangskonfiguration** sind voreingestellt.
HMÜ Bing	The most important parameters of the input and output configuration are preset.
Nach-KS	Die wichtigsten Parameter der **Eingangskonfiguration und Ausgangskonfiguration** sind voreingestellt.
HMÜ Bing	The most important parameters of the configuration of input and output configuration are preset.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.



SQ1: Ü ist ***nicht*** korrekt bzw. ***nicht*** klar dargestellt, d. h. nicht orthografisch.
SQ2: Ü ist ***nicht*** ideal für die Absicht des Satzes, d. h. motiviert den Nutzer ***nicht*** zum Handeln, zieht ***nicht*** seine Aufmerksamkeit an usw.
SQ3: Ü klingt ***nicht*** natürlich bzw. ***nicht*** idiomatisch.
CQ1: Ü gibt die Informationen im Ausgangstext ***nicht*** exakt wieder.
CQ2: Ü ist ***nicht*** leicht zu verstehen, d. h. ***nicht*** gut formuliert bzw. dargestellt.

Abbildung 5.140: „Keine Wortteile wegla.“ – Vergleich der Qualitätskriterien

Tabelle 5.142: Beispiel 89

Vor-KS	Schützen Sie das Gerät vor **Tropf- und Spritzwasser**.
HMÜ Systran	Protect the device from dripping and splash-water.
Nach-KS	Schützen Sie das Gerät vor **Tropfwasser und Spritzwasser**.
HMÜ Systran	Protect the device from dripping water and splash-water.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

In Tabelle 5.142 sanken nach der Regelanwendung die Stilqualität (– ,88 Punkte auf der Likert-Skala) und die Inhaltsqualität (– ,50 Punkte auf der Likert-Skala). Die Zeichensetzungsfehler (OR.1) in ‚splash-water' und der Wortklassenfehler (GR.7) in ‚splash' (anstelle von ‚splashing') wurden nach der Anwendung der Regel nicht behoben. Zudem fanden die Bewerter, dass die Wiederholung des Wortes ‚water' (nach KS) die Idiomatik der MÜ beeinträchtigte und ablenkend wirkte.

5.4.10.5.1 Korrelation zwischen den Fehlertypen und der Qualität

Auf Basis der Fehlerannotation zusammen mit der Humanevaluation gibt uns eine Spearman-Korrelationsanalyse Aufschluss, wie die Veränderung in der Fehleranzahl bei jedem Fehlertyp (Anz. nach KS – Anz. vor KS) mit den Qualitätsunterschieden (Q. nach KS – Q. vor KS) zusammenhängt. Grundsätzlich sind die meisten Korrelationen bei dieser Regel schwach. Nur bei der Stilqualität gab es zwei signifikante mittlere negative Korrelationen zwischen der Differenz in den Fehlertypen OR.1 „Zeichensetzungsfehler" und GR.10 „Falsche Wortstellung" einzeln und der Differenz in der Stilqualität. Die weiteren signifikanten Korrelationen zwischen der Differenz in den Fehlertypen OR.2 „Großschreibungsfehler" und SM.11 „Verwechslung des Sinns" einzeln und der Differenz in der Stilqualität waren schwache negative Korrelationen. (siehe Tabelle 5.143)

Ebenfalls waren die signifikanten Korrelationen zwischen der Differenz in den Fehlertypen LX.3 „Wort ausgelassen", LX.5 „Wort unübersetzt geblieben", GR.10 „Falsche Wortstellung", SM.11 „Verwechslung des Sinns" und SM.12 „Falsche Wahl" einzeln und der Differenz in der Inhaltsqualität schwache negative Korrelationen. (siehe Tabelle 5.143)

Weitere Korrelationen zwischen anderen einzelnen Fehlertypen und der Qualität konnten nicht nachgewiesen werden.

Tabelle 5.143: „Keine Wortteile wegla.“ – Korrelation zwischen den Fehlertypen und der Qualität

	N	p	ρ
Differenz SQ (nach KS − vor KS)			
Diff. der Anzahl der **OR.1 „Zeichensetzungsfehl.“**	87	,001	− ,338
Diff. der Anzahl der **OR.2 „Großschreibungsfehl.“**	87	,025	− ,240
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	87	,358	,100
Diff. der Anzahl der **LX.5 „W. unübers. geblieben“**	87	,992	,001
Diff. der Anzahl der **GR.10 „Falsche Wortstellung“**	87	,004	− ,310
Diff. der Anzahl der **SM.11 „Verwechsl. des Sinns“**	87	,016	− ,258
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	87	,720	,039
Differenz CQ (nach KS − vor KS)			
Diff. der Anzahl der **OR.1 „Zeichensetzungsfehl.“**	87	,126	− ,165
Diff. der Anzahl der **OR.2 „Großschreibungsfehl.“**	87	,089	− ,183
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	87	,029	− ,234
Diff. der Anzahl der **LX.5 „W. unübers. geblieben“**	87	,045	− ,216
Diff. der Anzahl der **GR.10 „Falsche Wortstellung“**	87	,008	− ,282
Diff. der Anzahl der **SM.11 „Verwechsl. des Sinns“**	87	,009	− ,278
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	87	,012	− ,267
Differenz allg. Q (nach KS − vor KS)			
Diff. der Anzahl der **OR.1 „Zeichensetzungsfehl.“**	87	,002	− ,325
Diff. der Anzahl der **OR.2 „Großschreibungsfehl.“**	87	,028	− ,236
Diff. der Anzahl der **LX.3 „Wort ausgelassen“**	87	,086	− ,185
Diff. der Anzahl der **LX.5 „W. unübers. geblieben“**	87	,165	− ,150
Diff. der Anzahl der **GR.10 „Falsche Wortstellung“**	87	,004	− ,308
Diff. der Anzahl der **SM.11 „Verwechsl. des Sinns“**	87	,001	− ,348
Diff. der Anzahl der **SM.12 „Falsche Wahl“**	87	,152	− ,155

*In der Tabelle werden nur die Fehlertypen dargestellt, die mindestens mit einer Qualitätsvariable signifikant korrelieren.

p: Signifikanz | nicht signifikant (p ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Bei dieser Regel fiel die Differenz in der Fehleranzahl bei den meisten Fehlertypen gering aus. Nur beim orthografischen Fehlertyp OR.1 „Zeichensetzung“ wurde eine signifikante Differenz registriert. Bei einer richtigen Zeichensetzung in Zusammenhang mit dieser Regel geht es um die richtige Verwendung des Bindestriches bei der Übersetzung aus dem Deutschen ins Englische aufgrund der orthografischen Unterschiede in den beiden Sprachen (siehe §4.5.2.3 „Diskussion der analysierten Regeln“). Eine Korrektur des Zeichensetzungsfehlers führte zur Verbesserung der Qualität. Allerdings kam dieser Fehler oft zusammen mit anderen Fehlertypen vor, sodass der positive Einfluss seiner Korrektur auf die MÜ-Qualität durch den anderen Fehlertyp im Endeffekt geschwächt wurde. In Tabelle 5.144 wurde der Zeichensetzungsfehler in ‚- cover buttons‘ nach KS behoben, allerdings blieb der semantische Fehler (in ‚cover buttons‘) unverändert.

Tabelle 5.144: Beispiel 90

Vor-KS	**Kunststoffgriffe und -deckelknöpfe** werden bei Verwendung im Backofen heiß.
HMÜ Systran	Plastic handles and - cover-buttons become hot when used in the oven.
Nach-KS	**Kunststoffgriffe und Kunststoffdeckelknöpfe** werden bei Verwendung im Backofen heiß.
HMÜ Systran	Plastic handles and plastic cover buttons become hot when used in the oven.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Daraufhin stieg die Stilqualität aufgrund der korrigierten orthografischen Darstellung um 0,25 Punkte und die Inhaltsqualität sank aufgrund der semantisch falschen Übersetzung um 0,38 Punkte auf der Likert-Skala.

5.4.10.5.2 Vergleich der Qualität auf Regel- und MÜ-Systemebene

Wie Abbildung 5.141 zeigt, sanken sowohl die Stilqualität als auch die Inhaltsqualität bei allen Systemen mit Ausnahme des HMÜ-Systems Systran, bei dem die Inhaltsqualität durchschnittlich leicht stieg.

Signifikante Rückgänge zeigten sich nur bei der Stilqualität und fanden bei drei MÜ-Systemen statt (Tabelle 5.145): bei dem NMÜ-System Google Translate

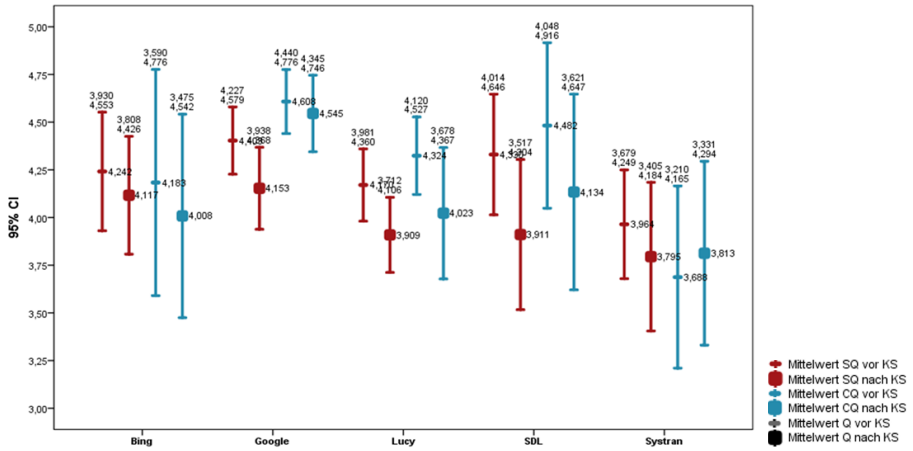


Abbildung 5.141: „Keine Wortteile wegla." – Mittelwerte der Qualität vor vs. nach KS bei den einzelnen MÜ-Systemen

(Diff_SQ – 2,9 %), dem RBMÜ-System Lucy (Diff_SQ – 5,7 %) und dem SMÜ-System SDL (Diff_SQ – 6,3 %). Bei den anderen Systemen war die Veränderung niedrig und entsprechend insignifikant. Für die Inhaltsqualität fiel die Differenz ebenfalls gering aus, sodass bei keinem System signifikante Werte verzeichnet wurden.

Tabelle 5.145: „Keine Wortteile wegla." – Signifikanz der Qualitätsveränderung bei den einzelnen MÜ-Systemen

	Differenz SQ (nach KS – vor KS)			Differenz CQ (nach KS – vor KS)			Differenz allg. Q (nach KS – vor KS)		
	N	p	z	N	p	z	N	p	z
Bing	15	,398	– ,844	15	,414	– ,818	15	,247	– 1,157
Google	22	,007	– 2,684	22	,139	– 1,479	22	,008	– 2,632
Lucy	22	,027	– 2,210	22	,264	– 1,116	22	,029	– 2,181
SDL	14	,010	– 2,567	14	,207	– 1,261	14	,069	– 1,819
Systran	14	,310	– 1,016	14	,925	– ,094	14	,285	– 1,069

p: Signifikanz z: Teststatistik nicht signifikant ($p \geq 0{,}05$)

Tabelle 5.146 zeigt, wie die Stil- und Inhaltsqualität durch die Regelanwendung unterschiedlich beeinflusst wurden.

Tabelle 5.146: Beispiel 91

Vor-KS	Im Falle der Auswahl der freien Konfiguration kann der **Start- und Endpunkt** frei gewählt werden.
HMÜ Systran	In case of choosing the free configuration, the starting and terminal can be selected freely.
Nach-KS	Im Falle der Auswahl der freien Konfiguration können der **Startpunkt und der Endpunkt** frei gewählt werden.
HMÜ Systran	In case of choosing the free configuration, the starting point and the terminal can be selected freely.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

In diesem Beispiel blieb der semantische Fehler in der Übersetzung von ‚Punkt' als ‚terminal' anstelle von ‚point' unverändert in beiden Szenarien. Die Qualitätswerte veränderten sich folgendermaßen: Die Stilqualität sank um 0,63 Punkte und die Inhaltsqualität stieg um 0,38 Punkte auf der Likert-Skala. Die Bewerterkommentare geben uns mehr Einblick in dieses Ergebnis; sie lauteten: „'starting point and terminal' is inconsistent construction"; „It sounds more natural to translate 'Endpunkt' as 'end point' since we already have 'starting point'. Then one can just use the word 'point' once. I suggest 'the starting and end points'". Somit ist der Rückgang in der Stilqualität durch die Inkonsistenz und die Redundanz begründet, wobei die MÜ aus Sicht der Bewerter nach KS inhaltlich ein wenig besser im Vergleich zu vor KS ausfiel.

5.4.10.5.3 Korrelation zwischen den Fehlertypen und der Qualität auf Regel- und MÜ-Systemebene

Anhand der Spearman-Korrelationsanalyse erwiesen sich bei drei MÜ-Systemen einige signifikante negative Korrelationen (Tabelle 5.147): Bei keiner weiteren KS-Regel zeigte das NMÜ-System Google Translate einen signifikanten Zusammenhang außer bei dieser Regel; hierbei gab es eine signifikante mittlere negative Korrelation zwischen der Differenz in Fehlertyp OR.1 „Zeichensetzungsfehler" und der Inhaltsqualität. Allerdings war, wie Tabelle 5.147 zeigt, das Signifikanzniveau sehr schwach (p = 0,049). Insgesamt beinhalteten zwei Sätze Zeichenset-

zungsfehler (OR.1), bei diesen kamen insgesamt 3 Fehler vor, die nach der Anwendung der KS korrigiert wurden. Aufgrund dieser sehr niedrigen Anzahl der Fälle (2 von den 24 bewerteten Sätzen) ist die Bedeutung der vorgeführten Korrelation gering.

Tabelle 5.147: „Keine Wortteile wegla.“ – Korrelationen zwischen den Fehlertypen und der Qualität bei den einzelnen MÜ-Systemen

	Google			Lucy			SDL		
	N	p	ρ	N	p	ρ	N	p	ρ
Differenz der Anzahl SQ (nach KS − vor KS)									
OR.1 „Zeichen.“	22	,314	− ,225				14	,034	− ,568
GR.10 „Wortst.“							14	,034	− ,568
SM.12 „f. Wahl“				22	,609	− ,116			
Differenz der Anzahl CQ (nach KS − vor KS)									
OR.1 „Zeichen.“	22	,049	− ,424				14	,090	− ,470
GR.10 „Wortst.“							14	,110	− ,446
SM.12 „f. Wahl“				22	,003	− ,597			
Differenz der Anzahl Q (nach KS − vor KS)									
OR.1 „Zeichen.“	22	,090	− ,370				14	,073	− ,493
GR.10 „Wortst.“							14	,090	− ,470
SM.12 „f. Wahl“				22	,010	− ,534			

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mind. einer Qualitätsvariable eine signifikante Korrelation aufweisen.

p: Signifikanz | nicht signifikant (ρ ≥ 0,05) | ρ: Korrelationskoeffizient

schwache Korrelation (ρ >=0,1) | mittlere Korrelation (ρ >= 0,3) | starke Korrelation (ρ >= 0,5)

Bei dem RBMÜ-System Lucy konnte nur eine signifikante starke negative Korrelation zwischen der Differenz in Fehlertyp SM.12 „Falsche Wahl“ und der Inhaltsqualität nachgewiesen werden. Bei dem SMÜ-System SDL erwies sich eine signifikante starke negative Korrelation zwischen der Differenz in den Fehlertypen OR.1 „Zeichensetzungsfehler und GR.10 „Wortstellungsfehler“ einzeln und der Stilqualität. (siehe Tabelle 5.147)

5.4.10.6 Vergleich der MÜ-Qualität vor vs. nach dem Weglassen von Wortteilen auf Annotationsgruppenebene

Die Qualitätsveränderung[63] der MÜ variierte in den verschiedenen Annotationsgruppen, nachdem die untersuchten Wörter vollständig ausgeschrieben wurden (nach KS) (Abbildung 5.142).

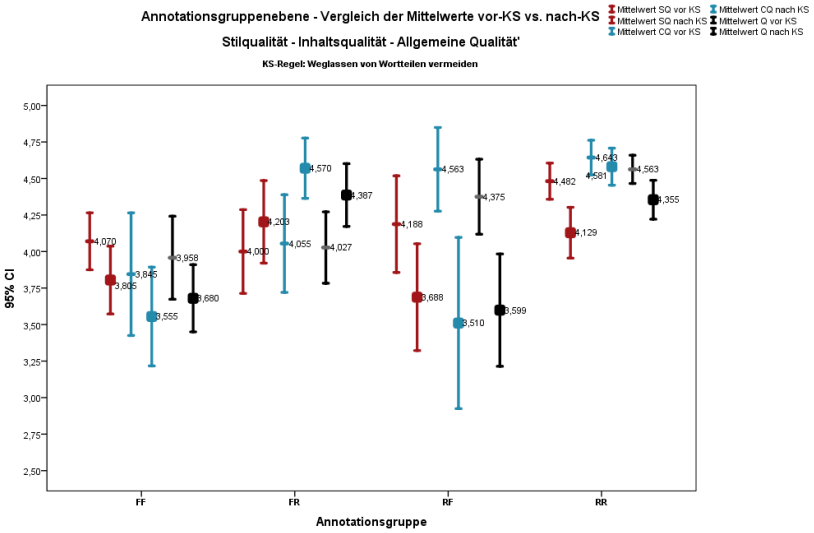


Abbildung 5.142: „Keine Wortteile wegla." – Mittelwerte der Qualität vor vs. nach KS auf Annotationsgruppenebene

In der Gruppe FF (Übersetzung vor und nach KS falsch) sanken die Stil- und Inhaltsqualität, wobei nur der Rückgang in der Stilqualität signifikant war (Tabelle 5.148). Die Bewerter fanden in vielen Fällen die MÜ bei der Formulierung in abgekürzter Form (vor KS) prägnanter und natürlicher.

In Tabelle 5.149 sanken die Stilqualität um 0,38 Punkte und die Inhaltsqualität um 0,25 auf der Likert-Skala. Bei der Formulierung in abgekürzter Form (vor KS) beeinflusste der semantische Übersetzungsfehler in ‚buttons' die Genauigkeit, die Verständlichkeit sowie die Idiomatik der MÜ. Auf der anderen Seite existierte nach der Anwendung der KS-Regel weiterhin der semantische Fehler und das Wort ‚plastic' wurde wiederholt. Diese Wiederholung fanden die Bewerter unnatürlich. So wurde die MÜ nach KS von den Bewertern folgendermaßen kommentiert: „I suggest deleting the second usage of 'plastic' because it's already

[63] Definitionen der Qualität unter §4.5.5.1.

Tabelle 5.148: „Keine Wortteile wegla." – Signifikanz der Qualitätsveränderung auf Annotationsgruppenebene

	N	p (Signifikanz)	Z (Teststatistik)
Annotationsgruppe FF			
Differenz SQ (nach KS − vor KS)	24	,012	− 2,521
Differenz CQ (nach KS − vor KS)	24	,177	− 1,351
Differenz allg. Q (nach KS − vor KS)	24	,018	− 2,361
Annotationsgruppe FR			
Differenz SQ (nach KS − vor KS)	16	,266	− 1,113
Differenz CQ (nach KS − vor KS)	16	,001	− 3,221
Differenz allg. Q (nach KS − vor KS)	16	,010	− 2,582
Annotationsgruppe RF			
Differenz SQ (nach KS − vor KS)	12	,028	− 2,201
Differenz CQ (nach KS − vor KS)	12	,004	− 2,875
Differenz allg. Q (nach KS − vor KS)	12	,004	− 2,904
Annotationsgruppe RR			
Differenz SQ (nach KS − vor KS)	35	< ,001	− 4,575
Differenz CQ (nach KS − vor KS)	35	,078	− 1,765
Differenz allg. Q (nach KS − vor KS)	35	< ,001	− 4,175

clear that the first time refers to everything that follows"; „'buttons' is something that can be pressed (Knopf). I suggest 'knobs' instead (something that can be used to lift or turn), so it should be 'Plastic handles and lid knobs'"

Erwartungsgemäß stiegen die Stil- und Inhaltsqualität in der Gruppe FR (MÜ falsch vor KS; richtig nach KS) und sanken in der Gruppe RF (MÜ richtig vor KS; falsch nach KS). In der Gruppe RF sanken die Stil- und Inhaltsqualität signifikant, nachdem die untersuchten Wörter vollständig ausgeschrieben wurden (nach KS), aufgrund der aufgetretenen Fehler im Vergleich zu der fehlerfreien Übersetzung bei der Formulierung mit Wortteilen (vor KS), siehe Tabelle 5.148. In der Gruppe FR war der Anstieg in der Stilqualität bei der Verwendung von vollständigen Wörtern (nach KS) insignifikant, während der Anstieg der Inhaltsqualität signifikant war (Tabelle 5.148). In Tabelle 5.150 verursachte die Verwendung von Wortteilen den semantischen Fehler in ‚is'. Dies beeinträchtigte deutlich die Verständ-

Tabelle 5.149: Beispiel 92

Vor-KS	**Kunststoffgriffe und -deckelknöpfe** werden bei Verwendung im Backofen heiß.
GNMÜ	Plastic handles and lid buttons become hot when used in the oven.
Nach-KS	**Kunststoffgriffe und Kunststoffdeckelknöpfe** werden bei Verwendung im Backofen heiß.
GNMÜ	Plastic handles and plastic cover buttons become hot when used in the oven.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

lichkeit der MÜ. Nachdem dieser Fehler bei der Formulierung mit vollständigen Wörtern (nach KS) behoben wurde, stieg die Inhaltsqualität um 1,25 und die Stilqualität um 0,25 auf der Likert-Skala.

Tabelle 5.150: Beispiel 93

Vor-KS	Die **Ist- und Sollwerte** des zweiten Regelkreises werden nach der Konfiguration angezeigt.
RBMÜ Lucy	The is- and required values of the second control loop will be displayed after configuration.
Nach-KS	Der **Istwert und der Sollwert** des zweiten Regelkreises werden nach der Konfiguration angezeigt.
RBMÜ Lucy	The actual value and the required value of the second control loop will be displayed after configuration.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Die Gruppe RR (Übersetzung vor und nach KS fehlerfrei) hatte den größten Anteil von 43 % der analysierten Sätze (siehe §5.4.10.3). In dieser Gruppe sanken sowohl die Stilqualität als auch die Inhaltsqualität, wobei nur der Rückgang in der Stilqualität signifikant war, siehe Tabelle 5.148. Die Humanevaluation zeigt, dass solange die Wortteile richtig übersetzt werden können und eine vollständige Formulierung des Worts für die Verständlichkeit nicht erforderlich ist, die

Verwendung von Wortteilen deutliche stilistische Vorteile mit sich bringt. In Tabelle 5.151 kritisierten die Bewerter nach der Regelanwendung die Redundanz in ‚errors'. Entsprechend sank die Stilqualität um 0,63 Punkte und die Inhaltsqualität um 0,13 Punkte auf der Likert-Skala.

Tabelle 5.151: Beispiel 94

Vor-KS	Innerhalb der Garantiezeit beseitigen wir alle Mängel des Gerätes, die auf **Material- oder Fabrikationsfehlern** beruhen.
GNMÜ	Within the guarantee period, we repair all device defects that are due to material or manufacturing errors.
Nach-KS	Innerhalb der Garantiezeit beseitigen wir alle Mängel des Gerätes, die auf **Materialfehlern oder Fabrikationsfehlern** beruhen.
GNMÜ	Within the guarantee period, we repair all device defects that are due to material errors or manufacturing errors.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Einer der wenigen Fälle (aus der Gruppe RR), in denen die Stil- und Inhaltsqualität nach der Regelanwendung stiegen (jeweils um 0,25 Punkte auf der Likert-Skala), demonstriert Tabelle 5.152.

Tabelle 5.152: Beispiel 95

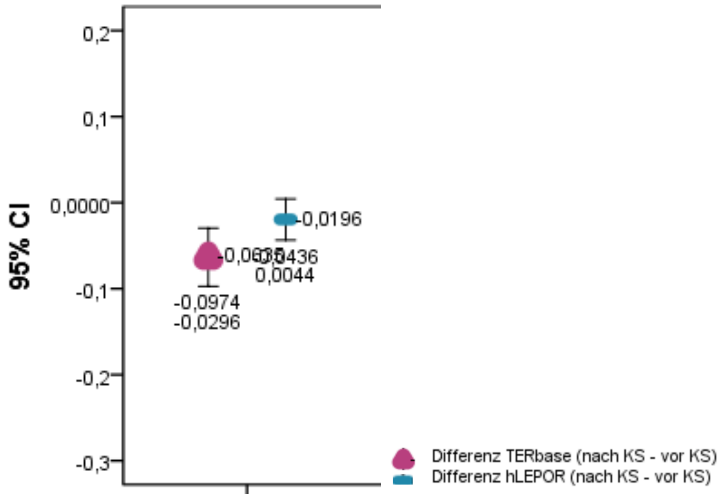
Vor-KS	Prüfen Sie, ob sich **Wasser–, Gasrohre** oder stromführende Leitungen im Bohrbereich befinden.
GNMÜ	Check whether there are water, gas pipes or power lines in the drilling area.
Nach-KS	Prüfen Sie, ob sich **Wasserrohre, Gasrohre** oder stromführende Leitungen im Bohrbereich befinden.
GNMÜ	Check whether there are water pipes, gas pipes or power lines in the drilling area.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Hierbei fanden die Bewerter die MÜ nach KS genauer und verständlicher im Vergleich zu der vor KS. Folglich beeinträchtigte die Wiederholung den Stil nicht.

5.4.10.7 Vergleich der AEM-Scores vor vs. nach dem Weglassen von Wortteilen sowie die Korrelation zwischen den AEM-Scores und der Qualität

Der Vergleich der AEM-Scores vor vs. nach dem Weglassen von Wortteilen zeigte sowohl mit TERbase als auch mit hLEPOR eine Verschlechterung der AEM-Scores (Abbildung 5.143).



Differenz = AEM-Score nach KS *minus* AEM-Score vor KS

Abbildung 5.143: „Keine Wortteile wegla." – Mittelwert der Differenz der AEM-Scores

Der Mittelwert der Differenz (nach KS – vor KS) im AEM-Score pro Satz lag für TERbase bei ,063 (SD = ,160) und für die hLEPOR bei ,020 (SD = ,113) mit einem 95%-Konfidenzintervall (Bootstrapping mit 1000 Stichproben). Die Differenzen (nach KS – vor KS) in TERbase und hLEPOR erwiesen sich als signifikant (z (N = 88) = – 3,777 / p = ,021) bzw. (z (N = 88) = – 1,843 / p = ,048). Dieses Ergebnis weist darauf hin, dass die Verwendung der vollständigen Form der Begriffe (nach KS) mit mehr Edits verbunden war.

5.4.10.7.1 Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Nach der Anwendung der KS-Regel sank die Stilqualität signifikant und die Inhaltsqualität nicht signifikant, siehe §5.4.10.5. Mithilfe des Spearman-Korrelationstests erwies sich ein signifikanter mittlerer positiver Zusammenhang zwi-

schen den Differenzen in den AEM-Scores von TERbase und hLEPOR und der Differenz in der allgemeinen Qualität. Bei der Formulierung der Begriffe in ihrer vollständigen Form verschlechterten sich die Scores der beiden AEMs und die Qualität sank.

Tabelle 5.153: „Keine Wortteile wegla." – Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	N	Signifikanz (p)	Korrelationskoeffizient (ρ)	Stärke der Korrelation
Korrelation zw. Differenz in der allg. Qualität und Differenz des TERbase-Scores (nach KS − vor KS)	87	< ,001	,465	mittlerer Zusammenhang
Korrelation zw. Differenz in der allg. Qualität und Differenz des hLEPOR-Scores (nach KS − vor KS)	87	< ,001	,454	mittlerer Zusammenhang

schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Nach diesem Ergebnis standen die Qualitätsveränderungen der Humanevaluation und der automatischen Evaluation in relativem Einklang, da der Qualitätsrückgang mit der Verschlechterung der AEM-Scores einherging.

5.4.10.8 Analyse der neunten Regel – Validierung der Hypothesen

Um die vorgestellten Ergebnisse auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der Ergebnisse der neunten analysierten Regel in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (−) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; *neg* steht für eine negative Korrelation und *pos* für eine positive Korrelation; <<>> steht für eine starke Korrelation und <> für eine mittlere Korrelation.[64]

[64] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Regel 9: Keine Wortteile weglassen

Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach dem Weglassen von Wortteilen

Fragestellung: Gibt es einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. **Anz.F. (–)**
Die Fehleranzahl sank leicht nach der Formulierung der Begriffe in vollständiger Form (nach KS).
Nach einer Aufteilung der untersuchten Begriffe nach Geläufigkeitsgrad zeigten abgekürzte Begriffe mit einer sehr niedrigen Geläufigkeit einen Rückgang der Fehleranzahl nach KS. Jedoch – aufgrund der kleinen Anzahl der Fälle – bedarf eine Analyse dieser Regel in Zusammenhang mit dem Geläufigkeitsgrad weiterer Forschung.

Auf Regel- und MÜ-Systemebene:
Alle Veränderungen in der Fehleranzahl nach der Regelanwendung waren nicht signifikant: ein Anstieg bei Bing, ein Rückgang bei Google, Lucy und Systran sowie gar keine Veränderung bei SDL.

Bi (+)
Go (–)
Lu (–)
Sy (–)
SD (=)

Zweiter Analysefaktor

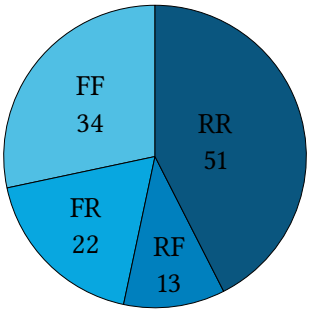


Abbildung 5.144: Aufteilung der Annotationsgruppen auf Regelebene

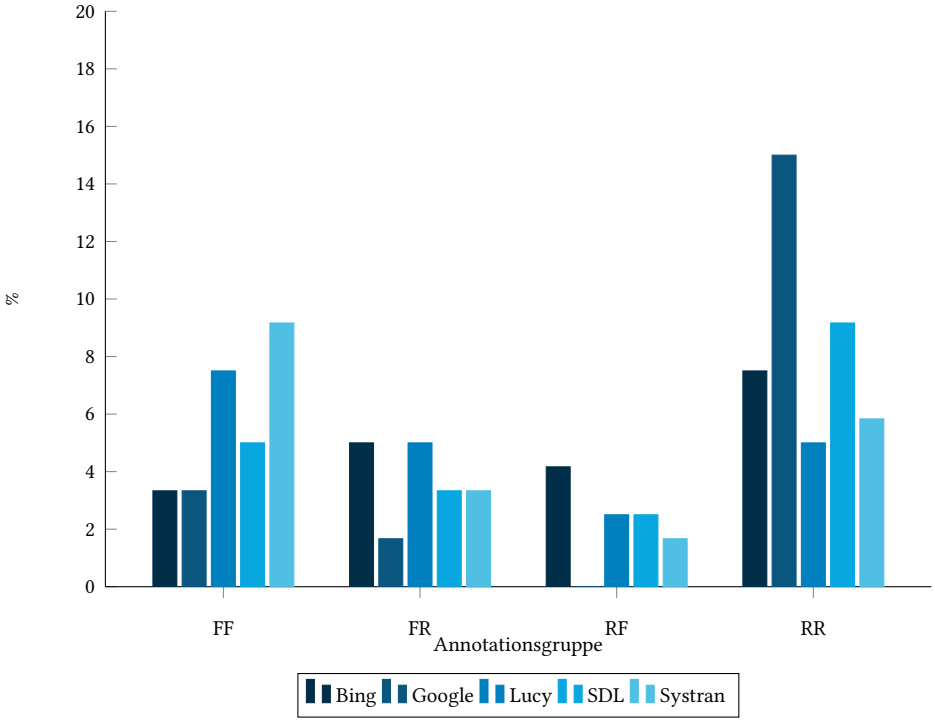


Abbildung 5.145: Aufteilung der Annotationsgruppen auf Regel- und MÜ-Systemebene

Dritter Analysefaktor: Vergleich der Fehlertypen vor vs. nach dem Weglassen von Wortteilen

Fragestellung: Beinhaltet die MÜ bestimmte Fehlertypen vor bzw. nach der Anwendung der KS-Regel?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H1 wurde nur für einen Fehlertyp bestätigt.
Die Fehleranzahl von OR.1 „Zeichensetzung" sank signifikant nach der Verwendung der vollständigen Form der Begriffe (nach KS).
Auf Regel- und MÜ-Systemebene:
Es gab bei keinem der Systeme signifikante Veränderungen in den Fehlertypen. Die Fehleranzahl bei den einzelnen Fehlertypen fiel im Allgemeinen niedrig aus.

OR.1 (–)

Vierter Analysefaktor: Vergleich der MÜ-Qualität vor vs. nach dem Weglassen von Wortteilen

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

Auf Regelebene:
H1 wurde nur für die Stilqualität bestätigt.
Die Stilqualität sank signifikant nach der Verwendung der vollständigen Form der Begriffe (nach KS).

SQ (–)

Die Inhaltsqualität sank leicht (nicht signifikant) nach KS. **CQ (–)**
Unabhängig vom Geläufigkeitsgrad der abgekürzten Begriffe sanken die SQ und CQ aller analysierten Fälle nach KS. Aufgrund der kleinen Anzahl der Fälle bedarf jedoch eine Analyse dieser Regel in Zusammenhang mit dem Geläufigkeitsgrad weiterer Forschung.

Auf Regel- und MÜ-Systemebene:

Die Stilqualität sank bei Google, Lucy und SDL signifikant (nach KS). **SQ (–):**
Go
Lu
SD

Die Stilqualität sank ebenfalls bei Bing und Systran, aber nicht signifikant.

Die Inhaltsqualität sank bei allen Systemen leicht mit Ausnahme von Systran, bei dem die CQ leicht anstieg (nach KS).

Fünfter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz in der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS – vor KS) und der Differenz in der Stil- bzw. Inhaltsqualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz in der Fehleranzahl eines bestimmten Fehlertyps und der Differenz in der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat

Auf Regelebene:

H1 wurde für zwei Fehlertypen wie folgt bestätigt:
Es bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz der Fehleranzahl des OR.1 „Zeichensetzungsfehler“ und GR.10 „Wortstellungsfehler“ einzeln und der Stilqualität.

***neg* OR.1 <> SQ**
***neg* GR.10 <> SQ**

Auf Regel- und MÜ-Systemebene:

Bei Google bestand ein signifikanter mittlerer negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des LX.1 „Zeichensetzungsfehler“ und der Differenz in der Inhaltsqualität.

Go
***neg* OR.1 <> CQ**

Bei Lucy bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des SM.12 „Falsche Wahl“ und der Differenz in der Inhaltsqualität.

Lu
***neg* SM.12 <<>> CQ**

Bei SDL bestand ein signifikanter starker negativer Zusammenhang zwischen der Differenz in der Fehleranzahl des OR.1 „Zeichensetzungsfehler“ und GR.10 „Wortstellungsfehler“ einzeln und der Differenz in der Stilqualität.

SD
***neg* OR.1 <<>> SQ**
***neg* GR.10 <<>> SQ**

Alle weiteren Korrelationen waren nicht signifikant.

Sechster Analysefaktor: Vergleich der MÜ-Qualität vor vs. nach dem Weglassen von Wortteilen auf Annotationsgruppenebene

Fragestellung: Gibt es einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat

H1 wurde bei der Annotationsgruppe FF nur für die Stilqualität bestätigt:
Die Stilqualität sank signifikant bei der vollständigen Formulierung der Begriffe (nach KS).

SQ (–)
CQ (–)

Bei der Annotationsgruppe FR stiegen die Stil- und Inhaltsqualität, allerdings war nur der Anstieg der Inhaltsqualität signifikant (nach KS).

CQ (+)
SQ (+)

Bei der Annotationsgruppe RF sanken die Stil- und Inhaltsqualität signifikant.

SQ (–)
CQ (–)

Bei der Annotationsgruppe RR sanken die Stil- und Inhaltsqualität, allerdings war nur der Rückgang der Stilqualität signifikant (nach KS).

SQ (–)
CQ (–)

Siebter Analysefaktor: Vergleich der AEM-Scores vor vs. nach dem Weglassen von Wortteilen

Fragestellung: Gibt es einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regel im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regel.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Die AEM-Scores von TERbase und hLEPOR verschlechterten sich signifikant nach der Verwendung der vollständigen Form der Begriffe (nach KS).

TERbase (–)
hLEPOR (–)

Achter Analysefaktor: Korrelation zwischen den Differenzen in den AEM-Scores und der Qualität

Fragestellung: Besteht ein Zusammenhang zwischen der Differenz in den AEM-Scores von TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS – vor KS) und der Differenz in der allgemeinen Qualität (Qualität nach KS – vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz in den AEM-Scores und der Differenz in der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz in den AEM-Scores und der Differenz in der allgemeinen Qualität.

Resultat

H0 wurde abgelehnt und somit H1 bestätigt.
Es bestand ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz in den TERbase-Scores und der Differenz der allgemeinen Qualität sowie ein signifikanter mittlerer positiver Zusammenhang zwischen der Differenz in den hLEPOR-Scores und der Differenz in der allgemeinen Qualität.

***pos* TERbase <> Q**
***pos* hLEPOR <> Q**

5.4.11 Übersicht der Ergebnisse auf Regelebene

Tabelle 5.154 bietet eine Übersicht über die Ergebnisse auf Regelebene.

Auf Regelebene hatten nur vier Regeln einen eindeutigen positiven Einfluss auf den MÜ-Output („anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“, „fvg – Funktionsverbgefüge vermeiden“, „kos – Konditionalsätze mit ‚Wenn‘ einleiten“ und „per – Konstruktionen mit ‚sein + zu + Infinitiv‘ vermeiden“): Die Fehleranzahl sank, die Stil- und Inhaltsqualität stiegen und die AEM-Scores von TERbase und hLEPOR verbesserten sich nach der Regelanwendung. Die positive Wirkung auf die Fehleranzahl, die Qualitätswerte sowie die AEM-Scores war bei den vier Regeln – mit Ausnahme der Verbesserung der AEM-Scores bei der Regel „kos“ – statistisch signifikant. Die signifikanten positiven Korrelationen zwischen den Veränderungen der Qualitätswerte und denen der AEM-Scores bekräftigen die positive Wirkung dieser Regeln (starke Korrelationen im Falle der Regeln „fvg“ und „kos“; mittlere Korrelationen im Falle der Regeln „anz“ und „per“). Unterschiedliche Fehlertypen gingen nach der Regelanwendung signifikant zurück. Der Rückgang dieser Fehlertypen korrelierte mit der Verbesserung der Stil- bzw. Inhaltsqualität. Auf der anderen Seite zeigten drei Regeln einen negativen Effekt („pak – Partizipialkonstruktion vermeiden“, „pas – Passiv vermeiden“ und „wte – Keine Wortteile weglassen“), wobei die Ergebnisse nicht immer statistisch signifikant waren: Ein Anstieg der Fehleranzahl war nur bei der Regel „pak“ signifikant; die Stilqualität litt eindeutig bei allen drei Regeln, während der Rückgang der Inhaltsqualität nur bei der Regel „pas“ signifikant war; die Verschlechterung beider AEM-Scores war bei allen drei Regeln signifikant. Der negative Einfluss dieser Regeln konnte durch eine positive Korrelation zwischen den Veränderungen der Qualitätswerte und denen der AEM-Scores bestätigt werden. Auch hier waren die beeinflussten Fehlertypen, wie Tabelle 5.154 zeigt, unterschiedlich. Schließlich konnten die letzten zwei Regeln („nsp – Eindeutige pronominale Bezüge verwenden“ und „prä – Überflüssige Präfixe vermeiden“) keinen eindeutigen Effekt anzeigen: Nur bei der Regel „prä“ war der Rückgang der Fehleranzahl signifikant; ansonsten waren die Veränderungen in den Qualitätswerten und den AEM-Scores nach der Anwendung beider Regeln insignifikant.

Tabelle 5.154: Übersicht der Ergebnisse auf Regelebene

Regel	Fehler-	Fehler-	Qualität		Fehlertypen <> Qualität		AEM-Scores		AEM-Scores <> allg. Qualität	
	anzahl	typen			Stilqualität	Inhaltsqualität	TERbase	hLEPOR	TERbase	hLEPOR
1: anz	(–)	–OR –GR	+SQ	+CQ	*neg* GR.10 <<>> SQ	*neg* GR.10 <<>> CQ	+	+	*pos* TERbase <> Q	*pos* hLEPOR <> Q
2: fvg	(–)	–SM	+SQ	+CQ	*neg* LX.4 <> SQ *neg* SM.11 <> SQ	*neg* SM.12 <> CQ	+	+	*pos* TERbase <<>> Q	*pos* hLEPOR <<>> Q
3: kos	(–)	–LX –LX	+SQ	+CQ	*neg* LX.3 <<>> SQ *neg* GR.10 <> SQ	*neg* LX.3 <<>> CQ	+	+	*pos* TERbase <<>> Q	*pos* hLEPOR <<>> Q
4: nsp	(–)	+LX.6 –SM.11	–SQ	+CQ	*neg* GR.10 <> SQ *neg* SM.11 <> SQ	*neg* LX.6 <> CQ *neg* GR.10 <> CQ *neg* SM.11 <> CQ	(–)	(–)	*pos* TERbase <> Q	*pos* hLEPOR <> Q
5: pak	+	+OR –GR	–SQ	–CQ	*neg* LX.4 <> SQ	*neg* LX.3 <> CQ *neg* LX.4 <> CQ *neg* GR.10 <> CQ	(–)	(–)	*pos* TERbase <> Q	*pos* hLEPOR <> Q
6: pas	+		–SQ	–CQ	*neg* LX.4 <> SQ *neg* GR.10 <> SQ	*neg* GR.10 <> CQ *neg* SM.11 <> CQ *neg* SM.12 <> CQ	(–)	(–)	*pos* TERbase <<>> Q	*pos* hLEPOR <<>> Q
7: per	(–)	+LX –GR –GR –GR	+SQ	+CQ	*neg* GR.8 <<>> SQ *neg* LX.4 <> SQ *neg* GR.10 <> SQ	*neg* GR.8 <<>> CQ *neg* GR.10 <> CQ	+	+	*pos* TERbase <> Q	*pos* hLEPOR <> Q
8: prä	(–)	–LX.4	+SQ	+CQ	*neg* LX.4 <> SQ *neg* SM.11 <> SQ	*neg* LX.4 <> CQ *neg* SM.11 <> CQ	+	+	*pos* TERbase <> Q	*pos* hLEPOR <> Q
9: wte	(–)	–OR	–SQ	–CQ	*neg* OR.1 <> SQ *neg* GR.10 <> SQ		(–)	(–)	*pos* TERbase <> Q	*pos* hLEPOR <> Q

SQ: Stilqualität; CQ: Inhaltsqualität; Q: allg. Qualität; Signifikant ($p < 0,5$) ; Blank: nicht signifikant; <> mittlere Korrelation ($\rho >= 0,3$);

<<>> starke Korrelation ($\rho >= 0,5$); neg: negative Korrelation; pos: positive Korrelation

anz: Für zitierte Oberflächentexte gerade Anführungszeichen verwenden; fvg: Funktionsverbgefüge vermeiden; kos: Konditionalsätze mit ‚Wenn' einleiten; nsp: Eindeutige pronominale Bezüge verwenden; pak: Partizipial-konstruktionen vermeiden; pas: Passiv vermeiden; per: Konstruktionen mit „sein + zu + Infinitiv" vermeiden; prä: Überflüssige Präfixe vermeiden; wte: Keine Wortteile weglassen

OR.1: Zeichensetzung; OR.2: Großschreibung; LX.3: Wort ausgelassen; LX.4: Zusätzliches Wort eingefügt; LX.5: Wort unübersetzt geblieben (auf DE wiedergegeben); LX.6: Konsistenzfehler; GR.7: Falsche Wortart/Wortklasse; GR.8: Falsches Verb (Zeitform, Komposition, Person); GR.9: Kongruenzfehler (Agreement); GR.10: Falsche Wortstellung; SM.11: Verwechslung des Sinns; SM.12: Falsche Wahl; SM.13: Kollokationsfehler

5.5 Analyse auf MÜ-Systemebene

In diesem Kapitel werden die Studienergebnisse auf Systemebene regelübergreifend dargestellt. Ziel ist die fünf verschiedenen Systeme vor vs. nach der Anwendung der KS-Regeln zu vergleichen. Wie §3.5.1 zeigt, beschäftigten sich die vorherigen Studien im Bereich der KS mit den älteren MÜ-Ansätzen. Ein Beitrag dieser Studie ist der (erstmalige) Vergleich eines NMÜ-Systems mit vier Systemen der früheren Ansätze im Hinblick auf die Anwendung von KS-Regeln. Hierbei wird folgender Frage nachgegangen: Inwiefern ist die Anwendung von KS-Regeln nach der Einführung der NMÜ zum Zwecke der maschinellen Übersetzbarkeit erforderlich?

Ein Überblick der analysierten MÜ-Systeme siehe Tabelle 5.155.

Tabelle 5.155: Überblick der analysierten MÜ-Systeme

Hybride MÜ-Systeme	Bing von Microsoft, Systran
Statistisches MÜ-System	SDL Free Translation
Regelbasiertes MÜ-System	Lucy LT KWIK Translator von Lucy Software and Services GmbH
Neuronale Netze-basiertes MÜ-System	Google Translate

Wie unter §4.5.1 erläutert, wurden zwei Hybridsysteme analysiert, da sie unterschiedlich aufgebaut werden. Dies wurde auch durch die in vielen Fällen unterschiedlichen Outputs ersichtlich. Im Folgenden gibt uns der erste Abschnitt eine Übersicht über die Analysefaktoren, die zugrundeliegenden Fragestellungen und Hypothesen sowie die statistische Auswertung. Danach folgen die Ergebnisse der quantitativen und qualitativen Analyse auf Systemebene. Bezüge auf vorherige Studien werden nicht in diesem Unterkapitel, sondern im Rahmen der Diskussion in §6 vorgenommen.

5.5.1 Analysefaktoren

Der Vergleich des MÜ-Outputs vor vs. nach der Anwendung aller analysierten KS-Regeln auf Systemebene erfolgte nach den folgenden neun Analysefaktoren:

Erster Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln

- Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?
- Variablen: Summe der Fehler und Mittelwert der Fehleranzahl (von allen Fehlertypen) innerhalb der KS-Stelle; Variablentyp: ordinal
- Statistische Auswertung: Deskriptive Statistiken auf Basis der Fehlerannotation; Abbildungen: Balken
- Hypothesen:

  H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.

  H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Die Variablen sind ordinal.

Zweiter Analysefaktor: Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln außerhalb der KS-Stelle bei der Gruppe RR

- Fragestellung: Wurde die KS-Stelle *bei einem bestimmten MÜ-System* vor und nach der Anwendung der KS-Regeln korrekt übersetzt und stieg *gleichzeitig* die Fehleranzahl *außerhalb* der KS-Stelle nach der Anwendung der KS-Regeln?
- Variablen: Differenzen in der Fehleranzahl bei der Annotationsgruppe RR (RR: MÜ innerhalb der KS-Stelle ist vor und nach der Anwendung der KS-Regeln fehlerfrei); Variablentyp: ordinal.
- Statistische Auswertung: Häufigkeitstabelle auf Basis der Fehlerannotation

Dritter Analysefaktor: Aufteilung der Annotationsgruppen

- In der Studie werden die Ergebnisse der Fehlerannotation in vier Annotationsgruppen unterteilt: (1) RR: MÜ ist vor und nach der Anwendung der KS-Regel fehlerfrei; (2) FF: MÜ beinhaltet vor und nach der Anwendung der KS-Regel Fehler; (3) RF: MÜ ist nur vor der Anwendung der KS-Regel fehlerfrei; (4) FR: MÜ ist nur nach der Anwendung der KS-Regel fehlerfrei.
- Fragestellung: Wie hoch ist der Prozentsatz jeder Annotationsgruppe *bei den einzelnen MÜ-Systemen*?
- Statistische Auswertung: Häufigkeiten mit Bootstrapping[65] auf Basis der Fehlerannotation; Abbildungen: Balken

Vierter Analysefaktor: Vergleich der Fehlertypen vor vs. nach der Anwendung der KS-Regeln

- Es werden die 13 analysierten Fehlertypen vor vs. nach der Anwendung der KS-Regeln verglichen.
- Fragestellung: Kommen bestimmte Fehlertypen *bei einem bestimmten MÜ-System* vor bzw. nach der Anwendung der KS-Regeln vor?

  Davon wird bei jedem MÜ-System abgeleitet, (1) ob bestimmte Fehlertypen, die vor der Anwendung der KS-Regeln existierten, nach der Anwendung der KS-Regeln eliminiert bzw. reduziert wurden; (2) ob bestimmte Fehlertypen erst nach der Anwendung der KS-Regeln auftraten bzw. deutlich stiegen (im Vergleich zu vor der Anwendung der KS-Regeln).
- Variablen: Fehler existiert ja/nein; Fehlertyp: dichotom. Summe der Fehler sowie Mittelwert der Fehleranzahl innerhalb der KS-Stelle; Variablentyp: ordinal
- Statistische Auswertung: Kreuztabellen auf Basis der Fehlerannotation; Abbildungen: Fehlerbalken
- Hypothesen:

  H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

[65] Bootstrapping ist ein statistisches Verfahren zur Schätzung der Stichprobenverteilung eines Schätzers durch erneute Stichprobenerstellung mit Ersatz aus der ursprünglichen Stichprobe. Es wird als ein nützliches Verfahren zum Testen der Modellstabilität angesehen. (IBM o.D.)

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: McNemar; Begründung der Testauswahl: Mithilfe des McNemar-Tests können zwei verbundene dichotome Parameter verglichen werden, somit kann eine mögliche signifikante Veränderung bei einem Fehlertyp vor vs. nach der Anwendung der KS-Regeln identifiziert werden.

Fünfter Analysefaktor: Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln

- Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?
- Variablen: Mittelwert der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch
- Statistische Auswertung: Deskriptive Statistiken auf Basis der Humanevaluation; Abbildungen: Fehlerbalken
- Hypothesen:

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

Sechster Analysefaktor: Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln auf Annotationsgruppenebene

- Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

  Davon wird abgeleitet, (1) ob bei der Gruppe RR die Stil- bzw. Inhaltsqualität vor bzw. nach der Anwendung der KS-Regeln höher ist, obwohl die MÜ in beiden Fällen fehlerfrei ausfällt; (2) ob bei der Gruppe FF die Stil-

bzw. Inhaltsqualität vor bzw. nach der Anwendung der KS-Regeln höher ist, obwohl die MÜ in beiden Fällen Fehler beinhaltet; (3) ob bei der Gruppe RF die Stil- bzw. Inhaltsqualität nach der Anwendung der KS-Regeln stieg, obwohl die MÜ nach der Anwendung der KS-Regeln Fehler beinhaltet und davor fehlerfrei war; (4) ob bei der Gruppe FR die Stil- bzw. Inhaltsqualität nach der Anwendung der KS-Regeln sank, obwohl die MÜ nach der Anwendung der KS-Regeln fehlerfrei ist und davor Fehler beinhaltete.

- Variablen: Mittelwert der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala in jeder Annotationsgruppe; Variablentyp: metrisch
- Statistische Auswertung: Deskriptive Statistiken auf Basis der Humanevaluation unter Aufteilung der Daten nach den Annotationsgruppen; Abbildungen: Fehlerbalken
- Hypothesen:

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

Siebter Analysefaktor: Korrelation zwischen den Fehlertypen und der Qualität

- Fragestellung: Besteht *bei einem bestimmten MÜ-System* ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS *minus* Qualität vor KS)?
- Variablen: Differenz der Mittelwerte der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch. Differenz der Fehleranzahl der einzelnen Fehlertypen; Variablentyp: ordinal.
- Statistische Auswertung: Spearman-Korrelationstest
- Hypothesen:

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

- Signifikanztest: Spearman-Korrelationstest; Begründung der Testauswahl: Eine der Variablen ist ordinal. Zudem setzt Spearman keine Anforderung an die Verteilung und die Linearität voraus.

Achter Analysefaktor: Vergleich der AEM-Scores vor vs. nach der Anwendung der KS-Regeln

- Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?
- Variablen: Mittelwert der AEM-Scores[66] und Differenzen der AEM-Scores (AEM-Score nach KS *minus* AEM-Score vor KS); Variablentyp: metrisch
- Statistische Auswertung: Deskriptive Statistiken auf Basis der automatischen Evaluation; Abbildungen: Fehlerbalken für die Differenzen der AEM-Scores[67]
- Hypothesen:

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

- Signifikanztest: Wilcoxon; Begründung der Testauswahl: Nicht alle Qualitätswerte sind normalverteilt. Wilcoxon kann bei normalverteilten sowie nicht-normalverteilten Variablen verwendet werden.

[66] Bei jeder MÜ wurde der Mittelwert der AEM-Scores auf Basis von zwei Referenzübersetzungen ermittelt; für eine genaue Beschreibung des Verfahrens siehe §4.5.6.4.

[67] Bei der Auswertung werden nur die Differenzen der AEM-Scores (und nicht die Mittelwerte der AEM-Scores) verwendet. Der Grund dafür ist, dass die Bewerter die komplette MÜ editiert haben. Ihre Edits können daher Stellen außerhalb der KS-Stelle umfassen. Da aber die MÜ vor und nach KS außerhalb der KS-Stelle vereinheitlicht wurden, wird hier davon ausgegangen, dass wir durch die Verwendung der Differenz (AEM-Score nach KS minus AEM-Score vor KS) nur die Edits innerhalb der KS-Stelle betrachten; für eine detaillierte Erläuterung siehe §4.5.6.4.

Neunter Analysefaktor: Korrelation zwischen den AEM-Scores-Differenzen und der Qualitätsdifferenz

- Fragestellung: Besteht *bei einem bestimmten MÜ-System* ein Zusammenhang zwischen der Differenz der AEM-Scores in TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS *minus* Mittelwert der AEM-Scores vor KS) und der Differenz der allgemeinen Qualität[68] (Qualität nach KS *minus* Qualität vor KS)?
- Variablen: Differenz der Mittelwerte der AEM-Scores sowie Differenz der Mittelwerte der Qualitätspunktzahlen der acht Teilnehmer auf der Likert-Skala; Variablentyp: metrisch
- Statistische Auswertung: Spearman-Korrelationstest
- Hypothesen:

  H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

  H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

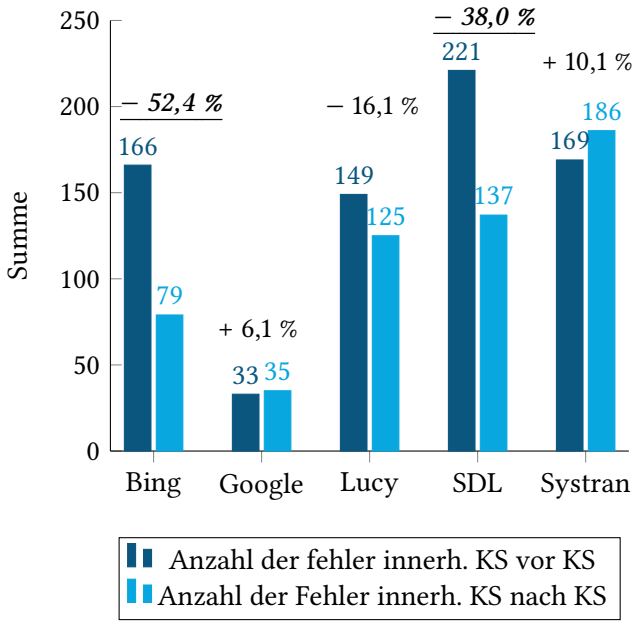
- Signifikanztest: Spearman-Korrelationstest; Begründung der Testauswahl: Nicht alle Variablen sind normalverteilt. Spearman setzt keine Anforderung an die Verteilung und die Linearität voraus.

5.5.2 Vergleich der Fehleranzahl vor vs. nach der Anwendung aller analysierten KS-Regeln

Die Veränderungsrichtung in der Fehleranzahl nach der Umsetzung der KS-Regeln war unterschiedlich:

Die *HMÜ*-Systeme reagierten unterschiedlich. Die Fehleranzahl sank nach der Anwendung der KS-Regeln bei Bing deutlich um 52 % (Abbildung 5.146). Die Differenz (nach KS *minus* vor KS) lag durchschnittlich bei − ,403 und erwies sich als signifikant (z (N = 216) = − 5,463 / p < ,001). Bei Systran hingegen gab es

[68] Die allgemeine Qualität ist der Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Untersuchung dieser Korrelation keine Unterscheidung zwischen der Stil- und Inhaltsqualität notwendig ist.



Signifikante Differenz vor vs. nach KS

Abbildung 5.146: Summe der Fehleranzahl vor vs. nach KS auf MÜ-Systemebene

einen kleinen (insignifikanten) Anstieg in der Fehleranzahl von 10,1 % mit einer durchschnittlichen Differenz von ,079 (Abbildung 5.147).[69]

Ebenfalls stieg die Fehleranzahl bei dem *NMÜ*-System Google Translate minimal um 6,1 %. Dies ist eine Differenz von nur zwei Fehlern (Abbildung 5.146). Zudem verzeichnete Google Translate die geringste Fehleranzahl sowohl vor als auch nach der Anwendung der KS-Regeln.

Bei dem *SMÜ*-System SDL sank die Fehleranzahl signifikant um 38 % (Abbildung 5.146). Die Differenz (nach KS minus vor KS) lag durchschnittlich bei – ,389 ($z (N = 216) = – 4{,}265 / p < ,001$).

Bei dem *RBMÜ*-System Lucy sank die Fehleranzahl insignifikant ($p = ,050$) um 16 % mit einer durchschnittlichen Differenz von – ,111 (Abbildung 5.147).

[69] Abbildung 5.147 ist wie folgt zu lesen: Die Punkte zeigen, wie hoch die durchschnittliche Fehleranzahl ausfällt (z. B. ,77 vor KS bei Bing). Die Fehlerbalken zeigen das realisierte 95 %-Konfidenzintervall (CI) für die durchschnittliche Fehleranzahl (in dem Fall beläuft sich das CI vor KS bei Bing auf ,65; ,89). Demnach würde die Fehleranzahl bei der Durchführung einer weiteren vergleichbaren Untersuchung mit einer Wahrscheinlichkeit von 95 % zwischen ,65 und ,89 liegen (vgl. Eckstein 2008: 81).

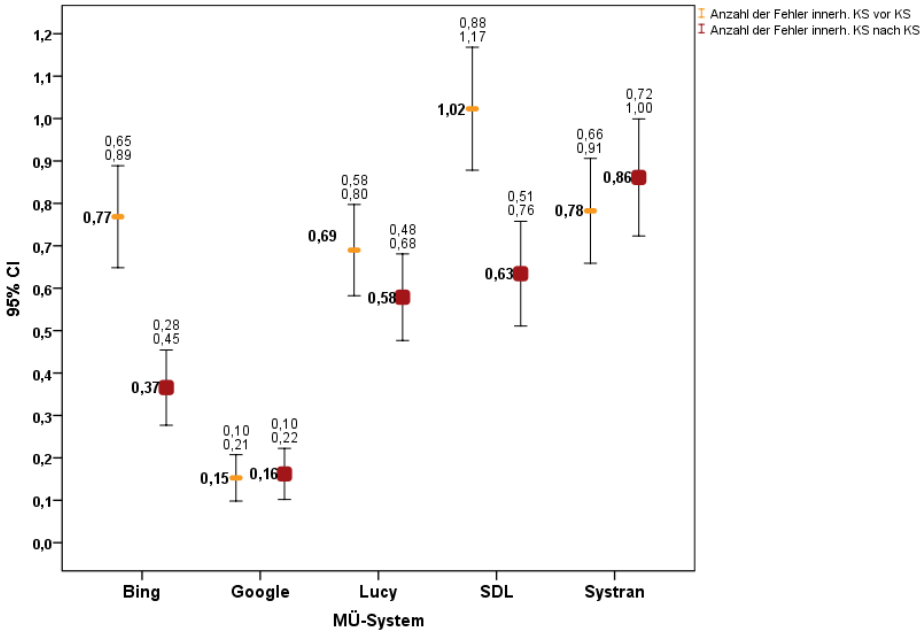


Abbildung 5.147: Mittelwert der Fehleranzahl pro Satz vor vs. nach KS auf MÜ-Systemebene

5.5.3 Vergleich der Fehleranzahl vor vs. nach KS außerhalb der KS-Stelle bei der Gruppe RR

Eine Untersuchung der Fehleranzahl außerhalb der KS-Stelle bei der Gruppe RR war erforderlich, um herauszufinden, ob bei einem bestimmten MÜ-System die KS-Stelle vor und nach der Anwendung der KS-Regeln korrekt übersetzt wurde und *gleichzeitig* die Fehleranzahl *außerhalb* der KS-Stelle nach der Anwendung der KS-Regeln stieg. In Tabelle 5.156 sind die Ergebnisse präsentiert (N = 490).

Tabelle 5.156 listet die Differenzen in der Fehleranzahl aller Übersetzungen auf, die innerhalb der KS-Stelle vor und nach der Anwendung der jeweiligen KS-Regel fehlerfrei waren, aber eine Veränderung in der Fehleranzahl außerhalb der KS-Stelle hatten. Aus den Ergebnissen geht hervor, dass kein deutlicher Anstieg der Fehleranzahl nach der Anwendung der KS bei einem bestimmten MÜ-System festgestellt werden konnte. Nur bei Google und Bing stieg die Fehleranzahl + 1 Fehler bei Google 10 Mal innerhalb von 183 Fällen und bei Bing 12 Mal innerhalb von 81 Fällen.

Tabelle 5.156: Häufigkeit der Differenz der Fehleranzahl außerhalb der KS-Stelle bei einer korrekten Übersetzung der KS-Stelle auf MÜ-Systemebene

Differenz Fehleranzahl nach KS − vor KS außerhalb KS	HMÜ Bing	GNMÜ	RBMÜ Lucy	SMÜ SDL	HMÜ Systran	Gesamt
+ 3	2	0	0	1	0	3
+ 2	1	0	0	1	0	2
+ 1	12	10	2	2	1	27
0	53	166	82	42	70	413
− 1	7	7	5	8	3	30
− 2	4	0	0	6	0	10
− 3	2	0	1	1	0	4
− 5	0	0	0	1	0	1
Gesamt	81	183	90	62	74	490

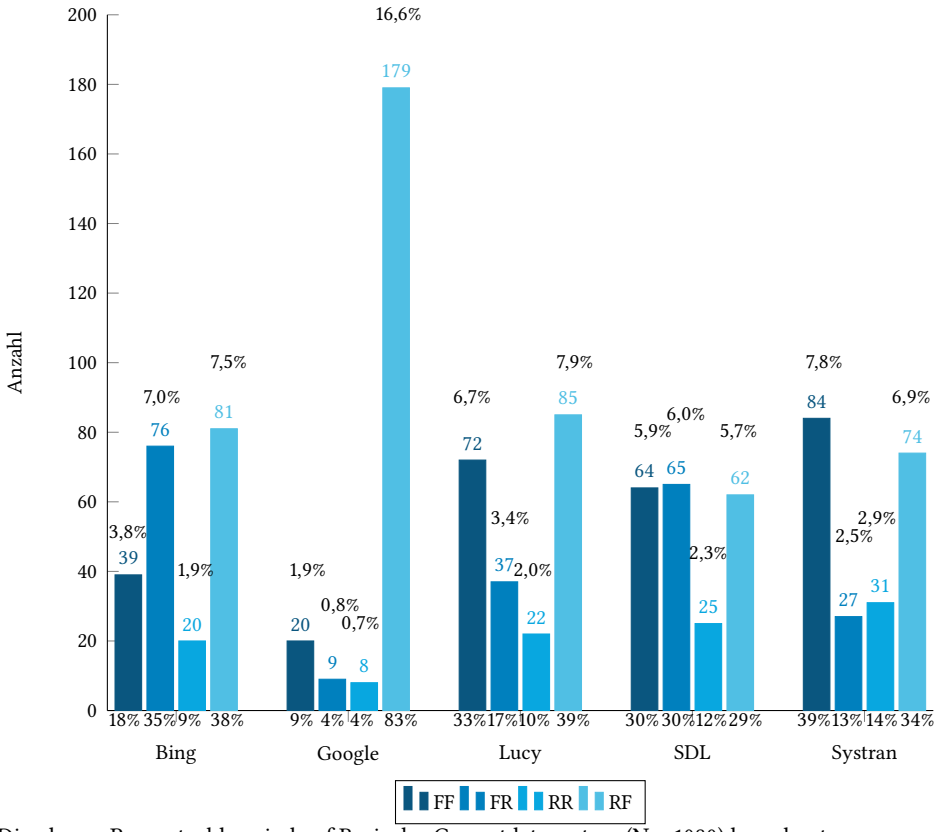
5.5.4 Aufteilung der Annotationsgruppen

Der Vergleich der Annotationsgruppen deckte bemerkungswerte Ergebnisse auf. Abbildung 5.148 verdeutlicht wie unterschiedlich die Verteilung der vier Annotationsgruppen in den drei älteren MÜ-Ansätzen (SMÜ, RBMÜ und HMÜ) auf einer Seite und dem neuronalen Ansatz auf der anderen Seite ausfällt.

Das *NMÜ*-System Google Translate produzierte die meisten Übersetzungen, die sowohl vor der Anwendung der KS-Regeln als auch nachher fehlerfrei (Gruppe RR) waren. Diese Gruppe betrug bei Google Translate 83 % seiner Übersetzungen (N = 216) bzw. knapp 17 % des Datensatzes (N = 1080). Mit 83 % RR war Google Translate in der Lage, mehr als doppelt so viel wie alle anderen Systeme mit und ohne die Anwendung der KS-Regeln fehlerfrei zu übersetzen, denn die Gruppe RR war in den anderen Systemen höchstens mit 39 % – beim *RBMÜ*-System Lucy – präsent (Abbildung 5.148).

Zudem war beim *RBMÜ*-System Lucy ein Drittel der Übersetzungen (33 % / N = 216) sowohl vor als auch nach der KS (Gruppe FF) falsch. Ferner konnten die Regeln in nur 17 % der Fälle Lucy dabei unterstützen, die vor-KS aufgetretenen Fehler zu beheben (Gruppe FR).

In den *HMÜ*-Systemen: Bing konnte hingegen mehr als ein Drittel (35 % / N = 216) der falschen Übersetzungen nach KS korrekt übersetzt werden (Gruppe FR). Im anderen HMÜ-System Systran war die Gruppe FF am meisten präsentiert (39 % / N = 216).



Die oberen Prozentzahlen sind auf Basis des Gesamtdatensatzes (N = 1080) berechnet.
Die unteren Prozentzahlen sind auf Systemebene (N = 216) berechnet.

Abbildung 5.148: Aufteilung der Annotationsgruppen auf MÜ-Systemebene

Im *SMÜ*-System SDL konnte etwas weniger als ein Drittel (30 % / N = 216) der falschen Übersetzungen nach KS korrekt übersetzt werden (Gruppe FR). Gleichzeitig gab es genauso viele (30 % / N = 216) Übersetzungen, die sowohl vor als auch nach der KS falsch übersetzt waren (Gruppe FF).

Die Ermittlung der Konfidenzintervalle (CI 95 %) der Aufteilung der Annotationsgruppen mithilfe eines Bootstrapping mit 1000 Stichproben ergab die Werte in Tabelle 5.157 (N = 216).

Das 95%-Konfidenzintervall besagt, dass bei der Durchführung einer weiteren vergleichbaren Untersuchung die Aufteilung mit einer Wahrscheinlichkeit von 95 % zwischen den aufgeführten Unter- und Oberwerten liegen würde. Deutlich

Tabelle 5.157: Konfidenzintervalle (CI 95 %) der Aufteilung der Annotationsgruppen auf MÜ-Systemebene

System	Annotationsgruppe	Häufigkeit	Prozente	Bootstrapping		
					95%-Konfidenzintervall	
				Verzerrung	Unterer Wert	Oberer Wert
HMÜ Bing	FF	39	18,1	−,1	13,0	23,3
	FR	76	35,2	,2	29,6	41,8
	RF	20	9,3	,0	5,4	13,6
	RR	81	37,5	−,1	30,8	43,8
	Gesamt	216	100,0	0,0	100,0	100,0
GNMÜ	FF	20	9,3	−,1	5,7	13,2
	FR	9	4,2	,0	1,8	7,2
	RF	8	3,7	,0	1,4	6,4
	RR	179	82,9	,0	78,1	87,8
	Gesamt	216	100,0	0,0	100,0	100,0
RBMÜ Lucy	FF	72	33,3	,0	27,2	39,9
	FR	37	17,1	,1	12,5	22,6
	RF	22	10,2	,0	6,3	14,5
	RR	85	39,4	−,1	32,8	45,9
	Gesamt	216	100,0	0,0	100,0	100,0
SMÜ SDL	FF	64	29,6	,0	23,6	36,0
	FR	65	30,1	,0	23,9	36,9
	RF	25	11,6	,0	7,5	16,1
	RR	62	28,7	−,1	22,7	35,2
	Gesamt	216	100,0	0,0	100,0	100,0
HMÜ Systran	FF	84	38,9	−,1	32,1	45,2
	FR	27	12,5	,0	8,1	17,2
	RF	31	14,4	,1	10,0	19,5
	RR	74	34,3	,0	28,0	40,6
	Gesamt	216	100,0	0,0	100,0	100,0

kleiner und entsprechend sicherer sind die Konfidenzintervalle bei dem NMÜ-System Google Translate im Vergleich zu den Konfidenzintervallen aller anderen Systemen.

5.5.5 Vergleich der Fehlertypen vor vs. nach der Anwendung aller analysierten KS-Regeln

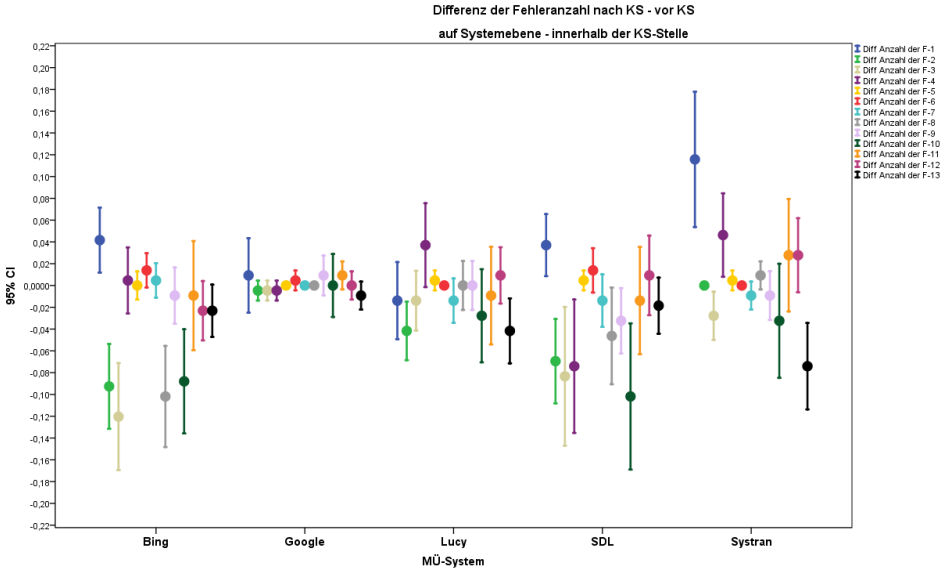
Die folgende Grafik veranschaulicht die Veränderungen in der Fehleranzahl, die nach der Anwendung der KS-Regeln bei den einzelnen Fehlertypen auftraten, und vergleicht in dieser Hinsicht die verschiedenen MÜ-Systeme. In allen Systemen mit Ausnahme von Google gab es mindestens zwei Fehlertypen, bei denen die Differenz in der Fehleranzahl (nach KS *minus* vor KS) signifikant war. Im Folgenden werden die signifikanten Fehlertypen bei den einzelnen Systemen erläutert.

Wie in der Analyse der Fehleranzahl (siehe §5.5.2) ersichtlich sank die Fehleranzahl bei dem HMÜ-System Bing und dem SMÜ-System SDL signifikant. Konkret signifikant war der Rückgang bei den folgenden Fehlertypen (Tabelle 5.158 und Tabelle 5.159).[70]

Tabelle 5.158: Bing – Fehlertypen mit signifikanter Veränderung nach der KS-Anwendung

	KS	N	M	CI 95%	SD	p
OR.1 „Zeichensetzung“	**vor**	216	,01	,00/,02	,096	,012
	nach		,05	,02/,08	,220	
OR.2 „Großschreibung“	**vor**	216	,10	,06/,14	,297	< ,001
	nach		,00	,00/,01	,068	
LX.3 „Wort ausgelassen“	**vor**	216	,16	,11/,21	,369	< ,001
	nach		,04	,01/,07	,200	
GR.8 „Falsches Verb“	**vor**	216	,13	,08/,17	,331	< ,001
	nach		,02	,00/,04	,151	
GR.10 „Wortstellungsfehler“	**vor**	216	,14	,09/,19	,364	,001
	nach		,06	,02/,09	,230	

[70] Die Tabellen in diesem Abschnitt demonstrieren nur die Fehlertypen, deren Unterschied signifikant ist.



Differenz der Fehleranzahl = Summe der Fehler nach KS *minus* Summe der Fehler vor KS
Fehlertyp 1: Orthografie – Zeichensetzung
Fehlertyp 2: Orthografie – Großschreibung
Fehlertyp 3: Lexik – Wort ausgelassen
Fehlertyp 4: Lexik – Wort zusätzlich falsch eingefügt
Fehlertyp 5: Lexik – Wort unübersetzt geblieben (auf DE wiedergegeben)
Fehlertyp 6: Lexik – Konsistenzfehler
Fehlertyp 7: Grammatik – Falsche Wortart / Wortklasse
Fehlertyp 8: Grammatik – Falsches Verb (Zeitform, Komposition, Person)
Fehlertyp 9: Grammatik – Kongruenzfehler (Agreement)
Fehlertyp 10: Grammatik – Falsche Wortstellung
Fehlertyp 11: Semantik – Verwechslung des Sinns
Fehlertyp 12: Semantik – Falsche Wahl
Fehlertyp 13: Semantik – Kollokationsfehler

Abbildung 5.149: Differenz der Fehleranzahl der einzelnen Fehlertypen auf MÜ-Systemebene

Tabelle 5.159: SDL – Fehlertypen mit signifikanter Veränderung nach der KS-Anwendung

	KS	N	M	CI 95%	SD	p
OR.1 „Zeichensetzung“	**vor**	216	,01	,00/,03	,117	,021
	nach		,05	,02/,08	,220	
OR.2 „Großschreibung“	**vor**	216	,09	,05/,13	,291	,001
	nach		,02	,00/,04	,151	
LX.3 „Wort ausgelassen“	**vor**	216	,17	,12/,22	,390	,011
	nach		,09	,05/,13	,300	
LX.4 „Zusätzliches Wort eingefügt“	**vor**	216	,13	,07/,18	,407	,036
	nach		,05	,02/,08	,241	
GR.9 „Kongruenzfehler“	**vor**	216	,05	,02/,08	,220	,004
	nach		,02	− ,01/,04	,192	
GR.10 „Wortstellungsfehler“	**vor**	216	,26	,19/,32	,480	,015
	nach		,16	,11/,21	,365	

Die Analyse auf Regel- und Systemebene zeigt, dass der Fehlertyp *OR.1 „Orthografie – Zeichensetzung“* bei den beiden *HMÜ*-Systemen und dem *SMÜ* SDL nur bei der Regel „Partizipialkonstruktion vermeiden“ signifikant stieg. In der Übersetzung nach der Anwendung der KS-Regel sollte ein Nebensatz mit ‚which‘ oder ‚that‘ gebildet werden (Tabelle 5.160).

Hierbei stellte das Relativpronomen für diese Systeme ein Ambiguitätsproblem dar. Das Relativpronomen wurde oft als Artikel interpretiert und dadurch entstand der Kommasetzungsfehler (in Tabelle 5.160 wurde ‚die‘ als ‚the‘ anstatt ‚which‘ übersetzt). Dementsprechend wurden z. B. bei dem *SMÜ*-System SDL 29 % der Sätze (7 von 24) in Bezug auf Fehlertyp OR.1 mit Partizipialkonstruktion (vor KS) richtig und nach KS falsch übersetzt.

Fehlertyp *OR.2 „Orthografie – Großschreibung“* sank nur in Zusammenhang mit Regel 1 „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ signifikant um 82 %. Dank der Verwendung von Anführungszeichen konnten das *SMÜ*-System SDL und das *HMÜ*-System Bing die Oberflächentexte als spezifische Begriffe bzw. Mehrwortentitäten erkennen und sie entsprechend großschreiben.

Tabelle 5.160: Beispiel 96

Vor-KS	Das Gerät verbindet sich mit **der neu gewählten Netzwerkadresse**.
SMÜ SDL	The device connects to the newly selected network address.
Nach-KS	Das Gerät verbindet sich mit **der Netzwerkadresse, die neu gewählt wird**.
SMÜ SDL	The device connects to the network address, the XXX newly selected.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Hochsignifikant war die Veränderung im Fehlertyp *LX.3 „Lexik – Wort ausgelassen“* bei der Regel „Konditionalsätze mit ‚Wenn‘ einleiten“: Alle 15 Fehler (N = 24), die vor der Verwendung der KS-Regel (bzw. bei einer Formulierung ohne ‚Wenn‘) auftreten, wurden nach der Verwendung von ‚Wenn‘ (nach-KS) eliminiert. Die Konstruktion einer Bedingung mit einem Verb am Satzbeginn (mit Ausnahme von ‚should‘) ist im Englischen nicht möglich. So stellen Verben am Satzbeginn von Konditionalsätzen die MÜ-Systeme vor ein Ambiguitätsproblem. Ein *RBMÜ*-System wie Lucy kann durch die Verwendung einer Systemregel diese Sprachunterschiede behandeln. In einem *SMÜ*-System wie SDL hingegen wäre es erforderlich, dass die Trainingsdaten diese Satzkonstruktion beinhalten, damit eine korrekte Übersetzung produziert werden kann. In Tabelle 5.161 konnte SDL den Konditionalsatz vor der Regalanwendung aufgrund des Fehlens der Konjunktion ‚Wenn‘ nicht als Bedingung erkennen.

Zudem wurde Fehlertyp *LX.3 „Lexik – Wort ausgelassen“* nach der Anwendung der KS-Regel 4 „Konstruktionen mit ‚sein + zu + Infinitiv‘ vermeiden“ korrigiert. Hierbei wurde das Verb (‚prüfen‘ in Tabelle 5.162) in der Passiversatzkonstruktion gar nicht übersetzt.

Die Trennung der Bestandteile der Passiversatzkonstruktion („sein“ am Anfang des Satzes und das Hauptverb des Satzes am Ende) bildete eine Parsing-Schwierigkeit für das *SMÜ*-System SDL und das *HMÜ*-System Bing, die zur Auslassung des Verbs führte (vor-KS in Tabelle 5.162). Die verwendete KS-Regel sieht vor, den Imperativ anstatt des Passiversatzes zu verwenden, wodurch das Verb geparst und übersetzt werden konnte (nach-KS in Tabelle 5.162).

Fehlertyp *LX.4 „Lexik – Wort zusätzlich falsch eingefügt“* war bei Regel 3 „Konditionalsätze mit ‚Wenn‘ einleiten“ in 5 der analysierten 24 Sätzen zu beobachten

Tabelle 5.161: Beispiel 97

Vor-KS	**Schließt** der Kontaktschalter, so wird der Raumdruck-Sollwert aktiv.
SMÜ SDL	XXX The contact switch closes, the room pressure setpoint becomes active.
Nach-KS	**Wenn** der Kontaktschalter **schließt**, wird der Raumdruck-Sollwert aktiv.
SMÜ SDL	If the contact switch closes, the room pressure setpoint becomes active.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Tabelle 5.162: Beispiel 98

Vor-KS	Die Teppichböden **sind** entsprechend den Liefer- und Zahlungsbedingungen **zu prüfen**.
SMÜ SDL	The carpets are XXX in accordance with the terms and conditions of delivery and payment.
Nach-KS	**Prüfen Sie** die Teppichböden entsprechend den Liefer- und Zahlungsbedingungen.
SMÜ SDL	Check the ***carpets*** in accordance with the terms and conditions of delivery and payment.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

(keine signifikanten Werte, aber beachtenswert). Alle 5 Sätze begannen die Bedingungsformulierung mit dem Verb ‚Ist', das vom *SMÜ*-System SDL als zusätzliches ‚is' übersetzt wurde (Tabelle 5.163).

Tabelle 5.163: Beispiel 99

Vor-KS	**Ist das Gerät oder das Netzkabel beschädigt, sofort den Netzstecker herausziehen.**
SMÜ SDL	XXX Is the ***appliance*** or the power cord is damaged, immediately ***disconnect*** the mains plug.
Nach-KS	**Prüfen Sie die Teppichböden entsprechend den Liefer- und Zahlungsbedingungen.**
SMÜ SDL	If the ***appliance*** or the power cord is damaged, immediately ***disconnect*** the mains plug.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens; XXX für ein fehlendes Wort oder Komma.

Fehlertyp *GR.9 „Grammatik – Kongruenzfehler"* trat in den Regeln vereinzelt auf und seine Veränderung vor und nach KS war bei keiner bestimmten Regel signifikant. Dies ist aber bei einem solchen grammatischen Fehler plausibel, denn ein Kongruenzfehler kann aus unterschiedlichen Gründen auftreten.

Die Verschachtelung einer Partizipialkonstruktion (Regel 5) war bei dem *SMÜ*-System SDL oft mit dem Fehlertyp *GR.10 „Grammatik – Falsche Wortstellung"* verbunden. Bei der Anwendung der KS-Regel wurde aus der Partizipialkonstruktion ein Nebensatz gebildet. Durch die Übersetzung des Nebensatzes mit SDL wurde der Wortstellungsfehler signifikant um knapp 69 % reduziert (vgl. Tabelle 5.164).

Ebenfalls sank Fehlertyp *GR.10 „Grammatik – Falsche Wortstellung"* bei Regel 7 „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden" signifikant um 91 %. Durch die Verwendung des Imperativs anstatt des Passiversatzes wurde der Satzbau für das *SMÜ*-System SDL vereinfacht. Dies verbesserte wiederum das Parsen und löst weitgehend die Wortstellungsproblematik.

Bei HMÜ-Systemen kann man keine eindeutige Aussage treffen, warum ein bestimmter Fehler auftrat oder eliminiert wurde. Das kann aufgrund der Regeln, die das System verwendet bzw. die im System fehlen oder der Trainingsdaten seiner statistischen Komponente erfolgen. In einer Black-Box-Analyse ist der Grund für das Auftreten eines Fehlers nicht ermittelbar. Bei dem *HMÜ*-System Systran war die Veränderung in der Fehleranzahl bei den folgenden Fehlertypen signifikant (Tabelle 5.165).

Tabelle 5.164: Beispiel 100

Vor-KS	**Die in der Betriebsanleitung angegebenen Fristen** für wiederkehrende Prüfungen sind einzuhalten.
SMÜ SDL	The specified in the operating instructions deadlines for periodic *tests must* be observed.
Nach-KS	**Die Fristen** für wiederkehrende Prüfungen, **die in der Betriebsanleitung angegeben sind**, sind einzuhalten.
SMÜ SDL	The deadlines for *periodic* tests specified in the Operator Manual *must* be observed.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Tabelle 5.165: Systran – Fehlertypen mit signifikanter Veränderung nach der KS-Anwendung

	KS	N	M	CI 95%	SD	p
OR.1 „Zeichensetzung“	**vor**	216	,03	,00/,06	,214	< ,001
	nach		,14	,08/,20	,455	
LX.3 „Wort ausgelassen“	**vor**	216	,04	,01/,07	,212	,031
	nach		,01	− ,01/,03	,136	
LX.4 „Zusätzliches Wort eingefügt“	**vor**	216	,01	,00/,03	,117	,035
	nach		,06	,03/,09	,257	
SM.13 „Kollokationsfehler“	**vor**	216	,11	,07/,16	,343	< ,001
	nach		,04	,01/,06	,189	

Bei dem *RBMÜ* Lucy war die Veränderung in der Fehleranzahl nur bei den folgenden zwei Fehlertypen signifikant (Tabelle 5.166).

Tabelle 5.166: Lucy – Fehlertypen mit signifikanter Veränderung nach der KS-Anwendung

	KS	N	M	CI 95%	SD	p
OR.2 „Großschreibung“	**vor**	216	,09	,05/,13	,291	,004
	nach		,05	,02/,08	,220	
SM.13 „Kollokationsfehler“	**vor**	216	,06	,02/,09	,230	,012
	nach		,01	,00/,03	,117	

Fehlertyp *OR.2 „Orthografie – Großschreibung“* sank nur in Zusammenhang mit Regel 1 „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ signifikant um 39 %. Dank der Verwendung von Anführungszeichen konnte das *RBMÜ*-System Lucy die Oberflächentexte als spezifische Begriffe bzw. Mehrwortentitäten erkennen und sie entsprechend großschreiben.

Fehlertyp *SM.13 „Semantik – Kollokationsfehler“* sank in Verbindung mit Regel 2 „Funktionsverbgefüge vermeiden“ signifikant um 90 %. Das *RBMÜ*-System Lucy hatte Schwierigkeiten Kollokationen (‚Schaden nehmen‘ vor-KS in Tabelle 5.167) korrekt zu übersetzen. Die Verwendung des bedeutungstragenden Verbs (‚beschädigt werden‘ nach-KS in Tabelle 5.167) vereinfacht die Satzsemantik und ermöglicht Lucy eine korrekte Übersetzung.

Tabelle 5.167: Beispiel 101

Vor-KS	Wird diese Regel nicht beachtet, kann der Motor **Schaden nehmen**.
RBMÜ Lucy	If this rule is not observed, the motor can take damage.
Nach-KS	Wird diese Regel nicht beachtet, kann der Motor **beschädigt werden**.
RBMÜ Lucy	If this rule is not observed, the motor can be damaged.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Bei dem *NMÜ*-System Google Translate gab es *keine* signifikanten Veränderungen in den Fehlertypen. Wie die Analyse der Fehleranzahl (unter §5.5.2) er-

wies, stieg die Fehleranzahl bei dem NMÜ-System Google Translate. Eine genaue Untersuchung der Regeln und der gestiegenen Fehlertypen bei diesem System zeigt, dass der Anstieg hauptsächlich bei Regel 5 „Partizipialkonstruktion vermeiden“ (+ 6 Fehler, wobei 4 davon je 2 Kommas in 2 Sätzen waren) und der Regel 4 „Eindeutige pronominale Bezüge verwenden“ (+ 4 Fehler) stattfand:

Das *NMÜ*-System Google Translate hat keine Schwierigkeiten, Partizipialkonstruktionen (Regel 5) zu übersetzen. Bei der Übersetzung des Nebensatzes (nach-KS in Tabelle 5.168) trat Fehlertyp *OR.1 „Orthografie – Zeichensetzung“* auf, konkret bei der Kommasetzung und der Auswahl des Relativpronomens ‚which‘ vs. ‚that‘. Für Letzteres zeigt die Humanevaluation, dass es sich um umstrittene Fälle handelte, in denen die Humanübersetzer Kontextinformationen benötigten.

Tabelle 5.168: Beispiel 102

Vor-KS	**Speziell auf diese Lautsprecher abgestimmtes Zubehör** erhalten Sie in unserem Webshop.
GNMÜ	Special accessories for these speakers are available in our webshop.
Nach-KS	**Zubehör, das speziell auf diese Lautsprecher abgestimmt ist**, erhalten Sie in unserem Webshop.
GNMÜ	Accessories, specially designed for these loudspeakers, are available in our webshop.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Der Anstieg der Fehleranzahl in Regel 4 „Eindeutige pronominale Bezüge verwenden“ (+ 4 Fehler) besteht aus 3 verschiedenen Fehlertypen: Fehlertyp LX.6 „Konsistenzfehler“ (+ 1), Fehlertyp GR.9 „Kongruenzfehler“ (+ 2) und Fehlertyp GR.10 „Wortstellungsfehler“ (+ 1) und hat somit keinen erkennbaren Grund bzw. kein rückführbares Muster.

Zudem stieg Fehlertyp *SM.11 „Semantik – Verwechslung des Sinns“* in Regel 3 „Konditionalsätze mit ‚Wenn‘ einleiten“ nach der Regelanwendung in *einem* Satz aufgrund von Ambiguität (‚wenn‘ wurde in Tabelle 5.169 als ‚when‘ anstatt ‚if‘ übersetzt). Auch hierfür würde – laut der Kommentare einiger Bewerter in der Humanevaluation – ein Humanübersetzer Kontextinformationen benötigen, um die Ambiguität klären zu können.

Ebenfalls trat Fehlertyp *SM.11 „Semantik – Verwechslung des Sinns“* in Regel 5 „Partizipialkonstruktion vermeiden“ nach der Anwendung der KS in *einem* Satz

Tabelle 5.169: Beispiel 103

Vor-KS	**Schließt** der Kontaktschalter, so wird der Raumdruck-Sollwert aktiv.
GNMÜ	If the contact switch closes, the room pressure setpoint becomes active.
Nach-KS	**Wenn** der Kontaktschalter **schließt**, wird der Raumdruck-Sollwert aktiv.
GNMÜ	When the contact switch closes, the room pressure setpoint becomes active.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

auf, in dem ‚which' anstatt ‚that' als Relativpronomen verwendet wurde (Tabelle 5.170).

Tabelle 5.170: Beispiel 104

Vor-KS	Das Gerät verbindet sich mit **der neu gewählten Netzwerkadresse**.
GNMÜ	The device connects to the newly selected network address.
Nach-KS	Das Gerät verbindet sich mit **der Netzwerkadresse**, **die neu gewählt wird**.
GNMÜ	The device connects to the network address, which is selected again.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Nun kann die alleinige Untersuchung der Fehleranzahl und ihrer Differenz keineswegs einen hinreichenden Hinweis auf die Qualität der Übersetzung geben und aufgrund der unterschiedlichen Gewichtung der verschiedenen Fehlertypen keine aussagekräftigen Ergebnisse liefern. Daher war es erforderlich, die Qualität auf einem anderen Weg zu beurteilen. In den folgenden Abschnitten werden die Ergebnisse der Qualitätsbewertung aus Sicht der Bewerter im Rahmen der Humanevaluation präsentiert und erläutert.

5.5.6 Vergleich der MÜ-Qualität vor vs. nach der Anwendung aller analysierten KS-Regeln

Ein Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln deckte Folgendes auf (Abbildung 5.150 und Abbildung 5.151): Während die Stil- und Inhaltsqualität bei allen älteren MÜ-Ansätzen nach der Anwendung der KS-Regeln stiegen, sanken beide bei dem NMÜ-System. Im Folgenden wird auf die Ergebnisse jedes Systems näher eingegangen.

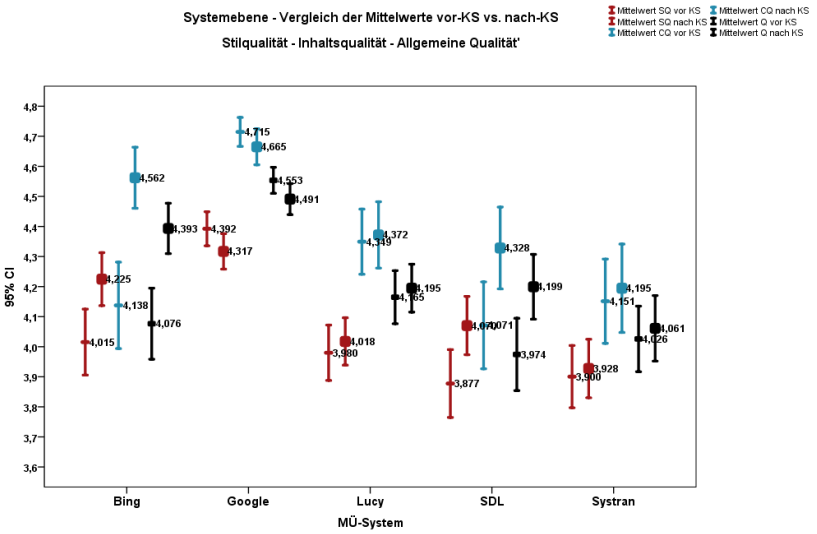
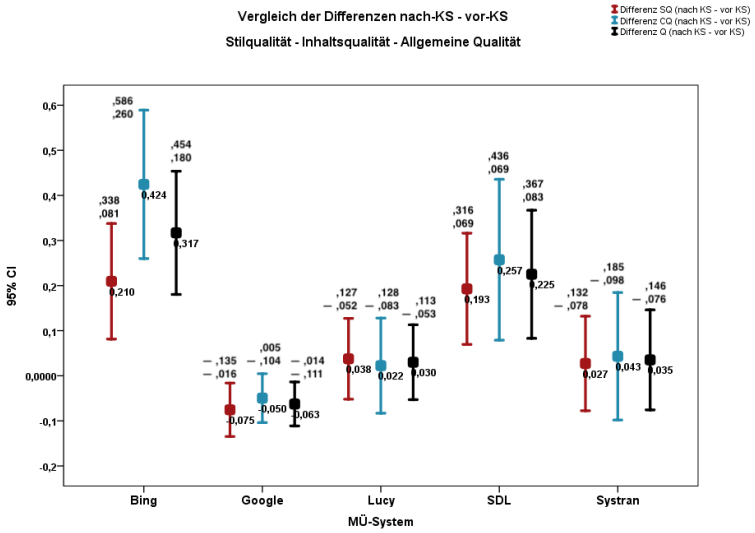


Abbildung 5.150: Mittelwerte der Qualität vor vs. nach KS auf MÜ-Systemebene

Beide *HMÜ*-Systeme Bing und Systran weisen einen Anstieg der Stil- und Inhaltsqualität auf, der Grad des Anstiegs war allerdings unterschiedlich (Abbildung 5.150 und Abbildung 5.151): Bei Bing war der Anstieg hochsignifikant und betrug für die Stilqualität + 5,23 % und für die Inhaltsqualität + 10,25 %. Die Differenzen (nach KS *minus* vor KS) in der Stil- und Inhaltsqualität erwiesen sich als signifikant (z (N = 139) = − 2,857 / p = ,004) bzw. (z (N = 139) = − 4,746 / p < ,001). Bei Systran zeigen die Ergebnisse einen kleinen Anstieg, der insignifikant war: für die Stilqualität + 0,72 % und für die Inhaltsqualität + 1,06 %.

Bei dem *SMÜ*-System SDL war der Anstieg der Stil- und Inhaltsqualität deutlich signifikant und betrug für die Stilqualität + 4,98 % mit einem Signifikanzniveau von (z (N = 153) = − 2,847 / p = ,004) und für die Inhaltsqualität + 6,31 % mit einem höheren Signifikanzniveau von (z (N = 153) = − 3,005 / p = ,003).



Qualitätsdifferenz = Qualitätswert nach KS − Qualitätswert vor KS

Abbildung 5.151: Mittelwert der Qualitätsdifferenzen auf MÜ-Systemebene

Im Falle des *RBMÜ*-Systems Lucy war der Anstieg sehr niedrig + 0,95 % für die Stilqualität bzw. + 0,53 % für die Inhaltsqualität und konnte somit keine signifikante Veränderung erweisen.

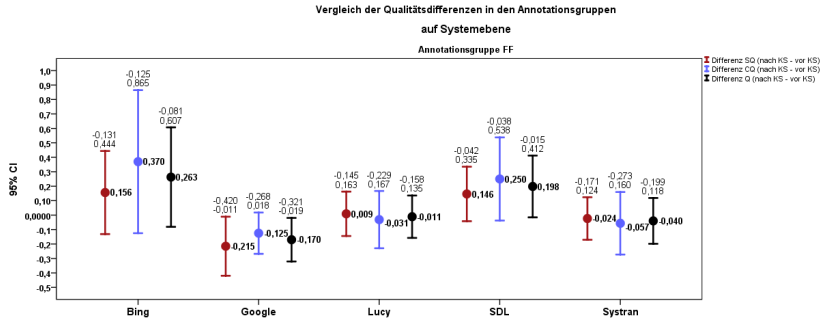
Das *NMÜ*-System Google Translate ist das einzige System, das einen Rückgang der Qualität verzeichnete: Die Stilqualität sank signifikant (− 1,71 % (z (N = 194) = − 2,338 / p = ,019). Hingegen sank die Inhaltsqualität um − 1,06 % und konnte keine signifikante Veränderung (z (N = 194) = − 1,481 / p = ,139) aufweisen.

Die Ergebnisse zeigen, dass die Anwendung der untersuchten Regeln für die vorherigen MÜ-Ansätze im Gegensatz zum NMÜ-Ansatz nützlich war. Wie die Analyse der Fehler zeigte, unterstützten die Regeln die RBMÜ-, SMÜ- und HMÜ-Systeme dabei, mehrere Fehlertypen zu reduzieren bzw. zu eliminieren. Die NMÜ-Architektur von Google Translate hingegen ermöglichte es, die Sätze vor und nach der Regelanwendung korrekt zu übersetzen. Ein genauer Einblick in die Bewertung der Teilnehmer zeigt, dass die Inhaltsqualität zweier korrekter MÜ vor und nach KS vergleichbar war. Gleichzeitig fanden die Teilnehmer die MÜ idiomatischer vor KS, was zum Rückgang der Stilqualität nach KS führte.

5.5.7 Vergleich der MÜ-Qualität vor vs. nach der Anwendung aller analysierten KS-Regeln auf Annotationsgruppenebene

Die Analyse der Qualitätsattribute auf Annotationsgruppenebene ist von Bedeutung, da sie die Qualitätsveränderung bei den einzelnen Annotationsgruppen klar darlegt. Während die Qualitätsveränderung bei den Gruppen FR und RF vorhersehbar sein sollte, nämlich ein Qualitätsanstieg bei der Gruppe FR bzw. ein Qualitätsrückgang bei der Gruppe RF, ist die Ermittlung der Qualitätsveränderung bei den Gruppen RR und FF von besonderem Interesse. Wie Abbildung 5.4 zeigt, waren die Gruppen RR und FF die größten Annotationsgruppen. Ein Vergleich des Qualitätsniveaus bei diesen dominanten Gruppen verrät, wie genau sich zwei vor- und nach-KS fehlerfreie Übersetzungen (Gruppe RR) bzw. zwei vor- und nach-KS fehlerhafte Übersetzungen (Gruppe FF) im Hinblick auf die Stil- und Inhaltsqualität unterscheiden. Das wiederum ermöglicht es, einen tieferen Einblick in den Einfluss der KS-Regeln zu gewinnen. Im Folgenden wird die Gegenüberstellung für jede Annotationsgruppe demonstriert:

Die *Gruppe FF* zeigt einen kleinen nicht signifikanten Anstieg der Qualität bei dem *HMÜ*-System Bing und dem *SMÜ*-System SDL. Bei dem *RBMÜ*-System Lucy und dem *HMÜ*-System Systran blieben die Qualitätswerte fast unverändert (Abbildung 5.152).



Qualitätsdifferenz = Qualitätswert nach KS − Qualitätswert vor KS

Abbildung 5.152: Gruppe FF – Qualitätsdifferenzen auf MÜ-Systemebene

Nur bei dem *NMÜ*-System Google Translate sanken beide Qualitätswerte bei der *Gruppe FF* signifikant (Abbildung 5.152): die Stilqualität um − 4,95 %, Mv = 4,340, Mn = 4,125, Mdiff = − ,215 (z (N = 18) = − 2,075 / *p* = *,038*) und die Inhaltsqualität um − 2,72 %, Mv = 4,590, Mn = 4,465, Mdiff = − ,125 (z (N = 18) = − 1,998

/ p = ,046). Die großen Abnahmen in der allgemeinen Qualität waren bei Regel 4 „Eindeutige pronominale Bezüge verwenden“ und Regel 5 „Partizipialkonstruktion vermeiden“ infolge des Rückgangs der Stilqualität. Die Wiederholung des Nomens bei Regel 4 (nach KS) wurde von den Teilnehmern als redundant betrachtet bzw. als unidiomatisch empfunden, was zu einem niedrigeren Score der Stilqualität führte (Genaueres dazu unter §5.4.5). Zudem hat Google Translate keine Schwierigkeit bei der Übersetzung von Partizipialkonstruktionen, siehe Tabelle 5.171. Obwohl der semantische Fehler in ‚time limits‘ in Tabelle 5.171 in beiden Szenarien unverändert blieb, wurde die Übersetzung der Partizipialkonstruktion (vor-KS) stilistisch höher bewertet.

Tabelle 5.171: Beispiel 105

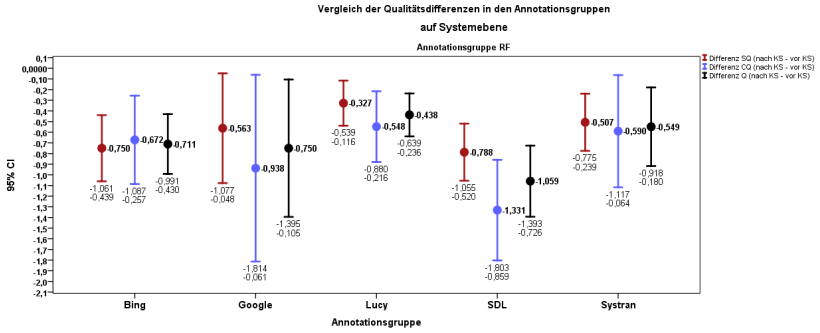
Vor-KS	**Die in der Betriebsanleitung angegebenen Fristen** für wiederkehrende Prüfungen sind einzuhalten.
GNMÜ	The time limits for ***periodic*** tests specified in the operating instructions must be observed.
Nach-KS	**Die Fristen** für wiederkehrende Prüfungen, **die in der Betriebsanleitung angegeben sind**, sind einzuhalten.
GNMÜ	The time limits for ***periodic*** tests, which are stated in the operating instructions, must be observed.

Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Erwartungsgemäß sank die Qualität in der *Gruppe RF* (Übersetzung vor KS richtig und nachher falsch) bei allen MÜ-Systemen.

Der Rückgang in der Qualität erwies sich bei allen Systemen mit Ausnahme der Inhaltsqualität bei Google Translate als signifikant (Tabelle 5.173). Insgesamt gab es 8 Übersetzungen (4 % der annotierten Übersetzungen) bei Google Translate, die in die Gruppe RF fielen (Abbildung 5.148); eine geringe Prozentzahl. Die 8 Übersetzungen wurden in der Humanevaluation bewertet. Das Ergebnis ist insbesondere bei orthografischen Fehlern gut vorstellbar. Wenn eine Übersetzung nach KS einen orthografischen Fehler enthält, bleibt sie – in vielen Fällen – inhaltlich qualitativ vergleichbar mit einer fehlerfreien Übersetzung vor der Anwendung der KS-Regeln. Diesen Fall demonstriert Tabelle 5.172, in dem die CQdiff 0 und die SQdiff − 0,25 betrugen.

Ebenfalls stieg die Qualität in der *Gruppe FR* (Übersetzung vor KS falsch und nachher richtig) erwartungsgemäß.



Qualitätsdifferenz = Qualitätswert nach KS − Qualitätswert vor KS

Abbildung 5.153: Gruppe RF – Qualitätsdifferenzen auf MÜ-Systemebene

Tabelle 5.172: Beispiel 106

Vor-KS	**Die in den Bedienungsanweisungen der eingebauten Geräte vorgeschriebenen Gebrauchsbedingungen** müssen strikt eingehalten werden.
GNMÜ	The operating conditions specified in the operating instructions of the installed devices must be strictly adhered to.
Nach-KS	**Die Gebrauchsbedingungen, die in den Bedienungsanweisungen der eingebauten Geräte vorgeschrieben sind**, müssen strikt eingehalten werden.
GNMÜ	The operating conditions, which are prescribed in the operating instructions of the installed devices, must be strictly adhered to.

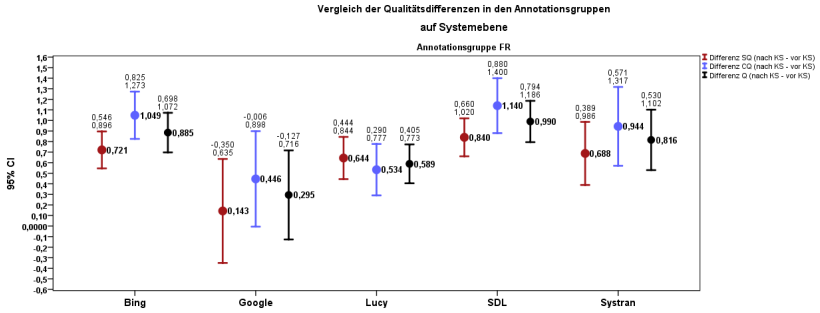
Die KS-Stelle ist **fett** dargestellt. Blau wird für die korrekten Tokens verwendet; Rot für die falschen Tokens.

Tabelle 5.173: Gruppe RF – Signifikanz der Qualitätsdifferenzen auf MÜ-Systemebene

	N	p (Signifikanz)	Z (Teststatistik)
Differenz SQ (nach KS *minus* vor KS)			
HMÜ Bing	16	,001	− 3,184
GNMÜ	8	,028	− 2,201
RBMÜ Lucy	21	,009	− 2,625
SMÜ SDL	20	< ,001	− 3,707
HMÜ Systran	18	,002	− 3,131
Differenz CQ (nach KS *minus* vor KS)			
HMÜ Bing	16	,001	− 3,176
GNMÜ	8	,063	− 1,859
RBMÜ Lucy	21	,001	− 3,387
SMÜ SDL	20	< ,001	− 3,706
HMÜ Systran	18	,035	− 2,106
Differenz allg. Q (nach KS *minus* vor KS)			
HMÜ Bing	16	,001	− 3,411
GNMÜ	8	,020	− 2,325
RBMÜ Lucy	21	< ,001	− 3,927
SMÜ SDL	20	< ,001	− 3,826
HMÜ Systran	18	,003	− 2,969

Der Rückgang in der Qualität erwies sich bei allen Systemen mit Ausnahme der Stilqualität bei Google Translate als signifikant (Tabelle 5.174). Es gab bei Google Translate nur 9 Übersetzungen (4 % der annotierten Übersetzungen), die in die Gruppe FR fielen (Abbildung 5.148). Ein sehr niedriger Prozentsatz, der darauf hinweist, dass das System selten falsch vor KS übersetzte und erst nach KS richtige Übersetzungen produzierte. Von den 9 Sätzen wurden nur 7 Sätze in der Humanevaluation bewertet.

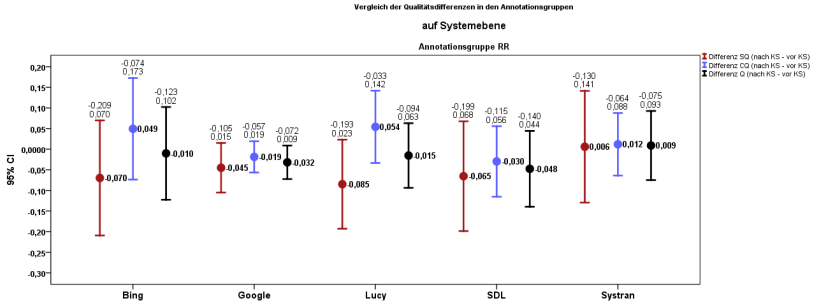
In der *Gruppe RR* (Übersetzung vor und nach KS richtig) waren die Qualitätsveränderungen gering und somit bei allen Systemen nicht signifikant. Obwohl die Ergebnisse statistisch nicht signifikant sind, geben uns die Qualitätsveränderungen bei der Gruppe RR einen Hinweis, wie die Anwendung der KS-Regeln die Qualitätsattribute beeinflusst hat. Abbildung 5.155 zeigt, dass bei dem Vergleich



Qualitätsdifferenz = Qualitätswert nach KS − Qualitätswert vor KS

Abbildung 5.154: Gruppe FR – Qualitätsdifferenzen auf MÜ-Systemebene

von zwei richtigen Übersetzungen vor und nach KS die Stilqualität nach KS bei fast allen Systemen sank. Nur bei dem HMÜ-System Systran stieg sie minimal (Mdiff = 0,006). Bei der Inhaltsqualität war das Ergebnis anders: Sie stieg bei dem HMÜ-System Bing, dem RBMÜ-System Lucy sowie dem HMÜ-System Systran und sank bei dem NMÜ-System Google Translate und dem SMÜ-System SDL.



Qualitätsdifferenz = Qualitätswert nach KS − Qualitätswert vor KS

Abbildung 5.155: Gruppe RR – Qualitätsdifferenzen auf MÜ-Systemebene

Im folgenden Abschnitt wird anhand der Berechnung der Korrelation zwischen der Differenz der Fehlertypen und den Qualitätsdifferenzen versucht, die hier ermittelten Qualitätsveränderungen näher zu beleuchten.

Tabelle 5.174: Gruppe FR – Signifikanz der Qualitätsdifferenzen auf MÜ-Systemebene

	N	p (Signifikanz)	Z (Teststatistik)
Differenz SQ (nach KS *minus* vor KS)			
HMÜ Bing	56	< ,001	− 5,842
GNMÜ	7	,442	− ,769
RBMÜ Lucy	26	< ,001	− 4,217
SMÜ SDL	50	< ,001	− 5,794
HMÜ Systran	20	,001	− 3,344
Differenz CQ (nach KS *minus* vor KS)			
HMÜ Bing	56	< ,001	− 6,053
GNMÜ	7	,034	− 2,120
RBMÜ Lucy	26	< ,001	− 3,965
SMÜ SDL	50	< ,001	− 5,844
HMÜ Systran	20	< ,001	− 3,851
Differenz allg. Q (nach KS *minus* vor KS)			
HMÜ Bing	56	< ,001	− 6,102
GNMÜ	7	,115	− 1,577
RBMÜ Lucy	26	< ,001	− 4,321
SMÜ SDL	50	< ,001	− 5,924
HMÜ Systran	20	< ,001	− 3,753

5.5.8 Korrelation zwischen der Differenz der Fehlertypen und den Qualitätsdifferenzen

Tabelle 5.175 – 5.177 zeigen die verschiedenen Korrelationen zwischen der Differenz der Fehlertypen und den Qualitätsdifferenzen bei den fünf MÜ-Systemen. In der linken Spalte sind die Korrelationen zwischen den Fehlertypen und SQ [1], CQ [2] und der allgemeinen Qualität[71] [3] aufgeführt. Der Korrelationskoeffizient einer signifikanten starken Korrelation (ρ >= 0,5) ist blau markiert. Mittlere

[71]Die allgemeine Qualität ist der Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Untersuchung dieser Korrelation keine Unterscheidung zwischen der Stil- und Inhaltsqualität notwendig ist.

Korrelationen (ρ >= 0,3) sind grün markiert. Schwache Korrelationen (ρ >=0,1) sind unmarkiert.

Das *HMÜ*-System Bing: Fehlertyp *OR.1 „Orthografie – Zeichensetzung"* korreliert mit einer signifikanten mittleren Korrelation mit der Stilqualität (ρ (N = 139) = − ,376 / p < ,001). Die Veränderung bei diesem Fehler war bei Regel 5 „Partizipialkonstruktion vermeiden". Durch die Bildung eines Nebensatzes mit ‚which' und ‚that' war die Kommasetzung für Bing problematisch, daher stieg die Fehleranzahl 10-fach (siehe §5.4.6.4).

Mit den grammatischen Fehlern korrelierte Bing mit Fehlertyp *GR.8 „Falsches Verb (Zeitform, Komposition, Person)"* sowie mit Fehlertyp *GR.10 „Falsche Wortstellung"* (Tabelle 5.175): Fehlertyp GR.8 wurde bei Regel 7 „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden" nach der Anwendung der Regel vollständig eliminiert (vorher war er in 19 von 24 Sätzen präsent). Bing hatte eine Schwierigkeit mit der Übersetzung des Passiversatzes. Nach der Verwendung des Imperativs wurden alle Fehler behoben (siehe §5.4.8.4). Ebenfalls sank Fehlertyp GR.8 „Falsches Verb" (nicht signifikant) bei der Regel „Überflüssige Präfixe vermeiden". Es war für Bing einfacher, Verben ohne Präfixe zu übersetzen (siehe §5.4.9.4). Entsprechend bestand eine signifikante mittlere negative Korrelation zwischen Fehlertyp GR.8 und der Stilqualität (ρ (N = 139) = − ,419 / p < ,001) und der Inhaltsqualität (ρ (N = 139) = − ,312 / p < ,001).

Fehlertyp *GR.10 „Falsche Wortstellung"* sank in Bing bei Regel 1 „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" (siehe §5.4.2.4) und Regel 3 „Konditionalsätze mit ‚Wenn' einleiten" (siehe §5.4.4.4) signifikant: Die Verwendung von Anführungszeichen und die Formulierung von Bedingungen mit ‚Wenn' anstatt mit einem Verb waren für Bing sehr hilfreich, um die Wortstellung zu korrigieren. Daher besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp GR.8 und der Stilqualität (ρ (N = 139) = − ,408 / p < ,001) und der Inhaltsqualität (ρ (N = 139) = − ,368 / p < ,001).

Auf lexikalischer Ebene korreliert Fehlertyp *LX.3 „Wort ausgelassen"* mit einer signifikanten mittleren Korrelation mit der Inhaltsqualität (ρ (N = 139) = − ,397 / p < ,001). Es gab eine hochsignifikante Veränderung (− 91,3 %) bei diesem Fehlertyp bei Regel 3 „Konditionalsätze mit ‚Wenn' einleiten" (siehe §5.4.4.4). Durch die Formulierung von Bedingungen mit einem Verb vor KS hat Bing ‚If' bei der Übersetzung ausgelassen. Zudem sank Fehlertyp LX.3 bei Regel 9 „Keine Wortteile weglassen" nicht signifikant in verschiedenen Fällen (siehe §5.4.10.4).

Das *HMÜ*-System Systran: Fehlertyp *SM.13 „Semantik – Kollokationsfehler"* korreliert mit einer signifikanten mittleren Korrelation mit der Stilqualität (ρ (N =

Tabelle 5.175: Korrelationen zwischen den Fehlertypen und der Qualität auf MÜ-Systemebene

	HMÜ Bing			GNMÜ			RBMÜ Lucy			SMÜ SDL			HMÜ Systran		
	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ
[1] Differenz der Anzahl SQ (nach KS *minus* vor KS)															
OR.1	139	< ,001	− ,376							153	,001	− ,277			
OR.2	139	,010	− ,210				156	< ,001	− ,367						
LX.3	139	,016	− ,203				156	,628	,039	153	< ,001	− ,401			
LX.4							156	,001	− ,252	153	< ,001	− ,323	133	,009	− ,226
LX.5	139	,166	− ,118				156	,343	− ,076	153	,289	− ,086			
LX.6				194	,189	− ,095									
GR.7	139	,024	− ,192				156	,318	− ,080	153	,123	− ,125			
GR.8	139	< ,001	− ,419							153	< ,001	− ,395			
GR.10	139	< ,001	− ,408	194	< ,001	− ,325	156	,002	− ,241	153	< ,001	− ,528	133	,001	− ,291
SM.11	139	,291	− ,090	194	,291	− ,090	156	,012	− ,200	153	,048	− ,160	133	,001	− ,285
SM.12				194	,833	,015	156	,065	− ,148	153	,012	− ,203	133	,015	− ,211
SM.13							156	,049	− ,158				133	< ,001	− ,304

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mindestens einer Qualitätsvariable signifikante Korrelationen aufweisen.

p: Signifikanz; nicht signifikant (p > 0,05) ; ρ: Korrelationskoeffizient; schwache Korrelation (ρ >=0,1); mittlere Korrelation (ρ >= 0,3) ; starke Korrelation (ρ >= 0,5)

Tabelle 5.176: Korrelationen zwischen den Fehlertypen und der Qualität auf MÜ-Systemebene (Fortsetzung)

	HMÜ Bing			GNMÜ			RBMÜ Lucy			SMÜ SDL			HMÜ Systran		
	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ
[2] Differenz der Anzahl CQ (nach KS *minus* vor KS)															
OR.1	139	,015	− ,207							153	,012	− ,204			
OR.2	139	,487	− ,059				156	,016	− ,192						
LX.3	139	< ,001	− ,397				156	,017	− ,192	153	< ,001	− ,534			
LX.4							156	,052	− ,156	153	,001	− ,261	133	,013	− ,215
LX.5	139	,047	− ,169				156	,001	− ,260	153	,018	− ,192			
LX.6				194	,030	− ,156									
GR.7	139	,006	− ,230				156	,016	− ,193	153	,001	− ,257			
GR.8	139	< ,001	− ,312							153	< ,001	− ,343			
GR.10	139	< ,001	− ,368	194	< ,001	− ,577	156	,001	− ,266	153	< ,001	− ,420	133	< ,001	− ,398
SM.11	139	,018	− ,200	194	,018	− ,200	156	,021	− ,184	153	,069	− ,148	133	< ,001	− ,399
SM.12				194	,002	− ,216	156	< ,001	− ,556	153	< ,001	− ,322	133	< ,001	− ,391
SM.13							156	,352	,075				133	,002	− ,260

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mindestens einer Qualitätsvariable signifikante Korrelationen aufweisen.

p: Signifikanz; nicht signifikant (p > 0,05) ; ρ: Korrelationskoeffizient; schwache Korrelation (ρ >=0,1); mittlere Korrelation (ρ >= 0,3) ; starke Korrelation (ρ >= 0,5)

Tabelle 5.177: Korrelationen zwischen den Fehlertypen und der Qualität auf MÜ-Systemebene (Fortsetzung)

	HMÜ Bing			GNMÜ			RBMÜ Lucy			SMÜ SDL			HMÜ Systran		
	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ	N	p	ρ
[3] Differenz der Anzahl Q (nach KS *minus* vor KS)															
OR.1	139	< ,001	− ,301							153	,002	− ,248			
OR.2	139	,105	− ,138				156	< ,001	− ,320						
LX.3	139	< ,001	− ,335				156	,212	− ,101	153	< ,001	− ,510			
LX.4							156	,003	− ,235	153	< ,001	− ,304	133	,005	− ,244
LX.5	139	,065	− ,157				156	,010	− ,206	153	,051	− ,158			
LX.6				194	,045	− ,144									
GR.7	139	,007	− ,229				156	,039	− ,166	153	,007	− ,215			
GR.8	139	< ,001	− ,385							153	< ,001	− ,387			
GR.10	139	< ,001	− ,413	194	< ,001	− ,518	156	< ,001	− ,299	153	< ,001	− ,493	133	< ,001	− ,391
SM.11	139	,056	− ,163	194	,039	− ,148	156	,005	− ,225	153	,045	− ,162	133	< ,001	− ,− ,389
SM.12				194	,124	− ,111	156	< ,001	− ,433	153	< ,001	− ,290	133	< ,001	− ,349
SM.13							156	,642	− ,038				133	< ,001	− ,309

*In der Tabelle werden nur die Fehlertypen dargestellt, die bei mindestens einer Qualitätsvariable signifikante Korrelationen aufweisen.
p: Signifikanz; nicht signifikant (p > 0,05) ; ρ: Korrelationskoeffizient; schwache Korrelation (ρ >=0,1); mittlere Korrelation (ρ >= 0,3) ; starke Korrelation (ρ >= 0,5)

133) = − ,304 / p < ,001). Die Veränderung bei diesem Fehlertyp war bei Regel 2 „Funktionsverbgefüge vermeiden“ (siehe §5.4.3.4). Systran hatte Schwierigkeiten Kollokationen korrekt zu übersetzen. Die Verwendung des bedeutungstragenden Verbs vereinfacht die Satzsemantik und ermöglicht Systran eine korrekte Übersetzung (eine Korrektur von knapp 94 %).

Fehlertyp *GR.10 „Falsche Wortstellung“* wurde in Systran nach Anwendung der Regel 5 „Partizipialkonstruktion vermeiden“ vollständig eliminiert (§5.4.6.4): Durch das Zerlegen der Partizipialkonstruktion und den Einbau eines Nebensatzes konnte Systran Fehlertyp GR.10 korrigieren (8 Fehler vorher; 0 nachher). Dementsprechend besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp GR.10 und der Inhaltsqualität (ρ (N = 133) = − ,398 / p < ,001).

Auf semantischer Ebene stieg Fehlertyp *SM.11 „Verwechslung des Sinns“* nach der Anwendung von Regel 6 „Passiv vermeiden“. Durch die Konvertierung des Passivs ins Aktiv (Bsp. „das Modul kann exportiert werden“ → „Sie können das Modul exportieren“) wurde das Subjekt ‚Sie‘ von Systran als ‚The‘ anstatt ‚You‘ übersetzt (siehe §5.4.7.4). Daher gibt es eine signifikante mittlere negative Korrelation zwischen Fehler SM.11 und der Inhaltsqualität (ρ (N = 133) = − ,399 / p < ,001). Ebenfalls besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp *SM.12 „Semantik – Falsche Wahl“* und der Inhaltsqualität (ρ (N = 133) = − ,391 / p < ,001). Allerdings ist Fehlertyp SM.12 in mehreren Regeln unterschiedlich auffindbar.

Das *SMÜ*-System SDL: Auf lexikalischer Ebene korreliert Fehlertyp *LX.3 „Wort ausgelassen“* mit einer signifikanten mittleren Korrelation mit der Stilqualität (ρ (N = 153) = − ,401 / p < ,001) und einer signifikanten starken Korrelation mit der Inhaltsqualität (ρ (N = 153) = − ,534 / p < ,001). Alle Fehler, die bei Regel 3 „Konditionalsätze mit ‚Wenn‘ einleiten“ vor der Anwendung der Regel vorkamen, wurden nach der Anwendung der Regel eliminiert (15 Fehler vorher; 0 Fehler nachher). Vergleichbar mit dem HMÜ-System Bing hat das SMÜ-System SDL durch die Formulierung von Bedingungen mit einem Verb (vor KS) ‚If‘ bei der Übersetzung ausgelassen (siehe §5.4.4.4). Außerdem korreliert Fehlertyp *LX.4 „Lexik – Wort zusätzlich falsch eingefügt“* mit einer signifikanten mittleren Korrelation mit der Stilqualität (ρ (N = 153) = − ,323 / p < ,001). Fehlertyp LX.4 war bei mehreren Regeln vertreten.

Auf grammatischer Ebene sank Fehlertyp *GR.8 „Falsches Verb (Zeitform, Komposition, Person)“* vergleichbar mit dem HMÜ-System Bing bei Regel 7 „Konstruktionen mit ‚sein + zu + Infinitiv‘ vermeiden“ nach der Anwendung der Regel um knapp 86 % (siehe §5.4.8.4). Auch SDL hatte Schwierigkeiten mit der Übersetzung

des Passiversatzes. Nach der Anwendung der Regel wurde der Imperativ korrekt übersetzt. Entsprechend besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp GR.8 und der Stilqualität (ρ (N = 153) = − ,395 / p < ,001) und der Inhaltsqualität (ρ (N = 153) = − ,343 / p < ,001).

Fehlertyp *GR.10 „Falsche Wortstellung"* sank vergleichbar mit dem HMÜ-System Systran nach Anwendung der Regel 5 „Partizipialkonstruktion vermeiden" um knapp 69 % (siehe §5.4.6.4): Durch das Zerlegen der Partizipialkonstruktion und den Einbau eines Nebensatzes konnte SDL Fehlertyp GR.10 minimieren (16 Fehler vorher; 5 nachher). Zudem wurde Fehlertyp GR.10 bei Regel 7 „Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden" fast vollkommen eliminiert (11 Fehler vorher; 1 nachher). Auch hier wurde der Imperativ vom SMÜ-System SDL korrekt übersetzt (siehe §5.4.8.4). Dementsprechend besteht eine signifikante starke negative Korrelation zwischen Fehlertyp GR.10 und der Stilqualität (ρ (N = 153) = − ,528 / p < ,001) sowie eine signifikante mittlere negative Korrelation mit der Inhaltsqualität (ρ (N = 153) = − ,420 / p < ,001).

Auf semantischer Ebene besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp *SM.12 „Semantik – Falsche Wahl"* und der Inhaltsqualität (ρ (N = 153) = − ,322 / p < ,001). Dennoch ist Fehlertyp SM.12 in mehreren Regeln unterschiedlich auffindbar.

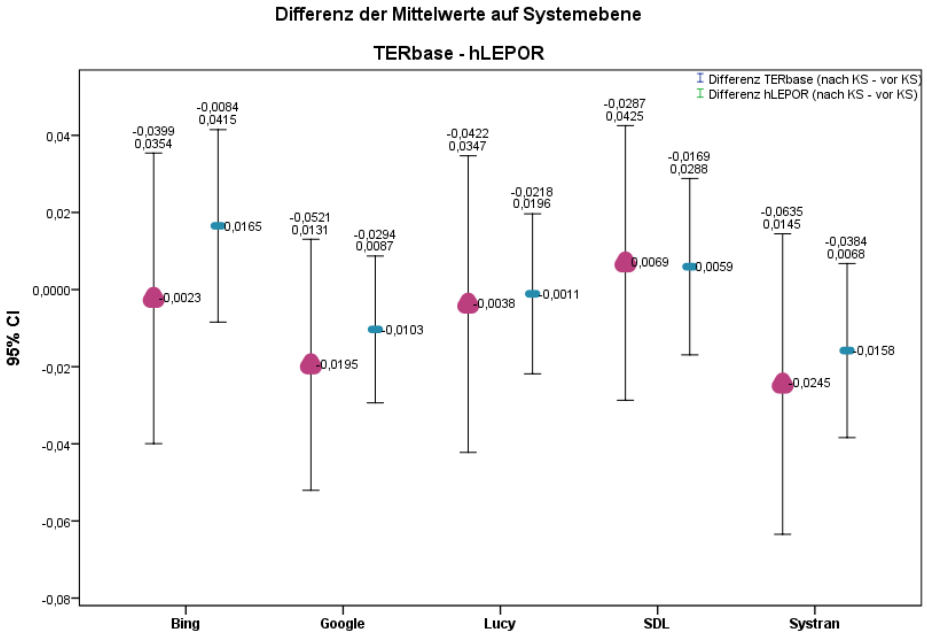
Das *RBMÜ*-System Lucy: Fehlertyp *OR.2 „Orthografie – Großschreibung"* sank in Lucy bei der Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" signifikant um knapp 39 % (siehe §5.4.2.4): die Verwendung von Anführungszeichen war für das RBMÜ-System Lucy hilfreich, um die Oberflächentexte als spezifische Begriffe bzw. Mehrwortentitäten zu erkennen und sie entsprechend großzuschreiben. Daher besteht eine signifikante mittlere negative Korrelation zwischen Fehlertyp OR.2 und der Stilqualität (ρ (N = 156) = − ,367 / p < ,001).

Vergleichbar mit dem SMÜ-System SDL wiederholte sich Fehlertyp *SM.12 „Semantik – Falsche Wahl"* in mehreren Regeln. Daher besteht eine signifikante starke negative Korrelation zwischen Fehlertyp SM.12 und der Inhaltsqualität (ρ (N = 156) = − ,556 / p < ,001).

Das *NMÜ*-System Google Translate: Der einzige Fehler, der mit beiden Qualitätsattributen korreliert, ist Fehlertyp *GR.10 „Grammatik – Falsche Wortstellung"*: eine signifikante mittlere negative Korrelation mit der Stilqualität (ρ (N = 194) = − ,325 / p < ,001) und eine signifikante starke negative Korrelation mit der Inhaltsqualität (ρ (N = 194) = − ,577 / p < ,001). Allerdings ist Fehlertyp GR.10 in mehreren Regeln unterschiedlich auffindbar.

5.5.9 Vergleich der AEM-Scores vor vs. nach der Anwendung aller analysierten KS-Regeln

Wie Abbildung 5.156 veranschaulicht, waren die Differenzen[72] der Mittelwerte der AEM-Scores sowohl mit TERbase als auch mit hLEPOR[73] bei allen MÜ-Systemen minimal und lagen zwischen + 0,017 und − 0,025. Es zeigte sich daher kein signifikanter Unterschied vor vs. nach der Anwendung der analysierten KS-Regeln.



Differenz = AEM-Score nach KS minus AEM-Score vor KS

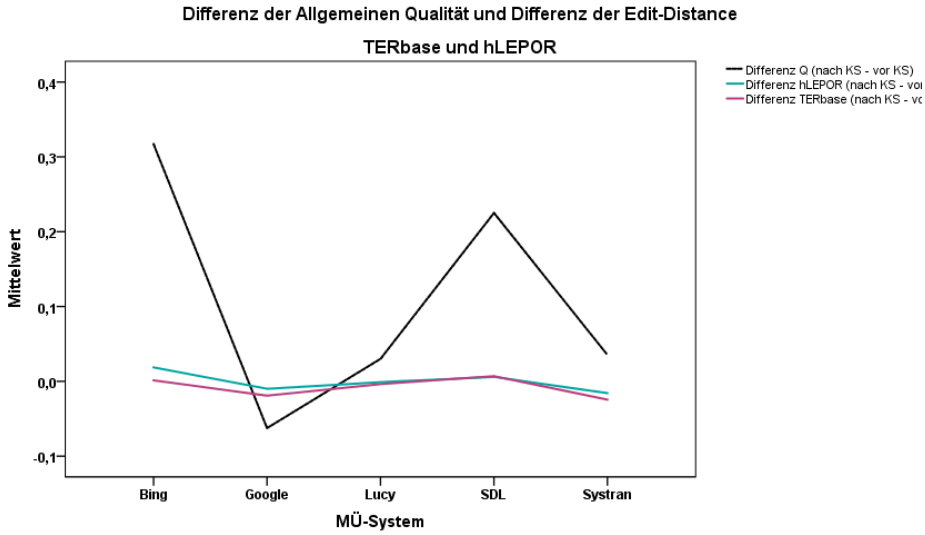
Abbildung 5.156: Mittelwert der Differenz des TERbase-Scores und hLEPOR-Scores auf MÜ-Systemebene

[72] Bei der Analyse der AEM-Scores wird nur die Differenz der Mittelwerte (nicht die Mittelwerte) berücksichtigt, da die Mittelwerte AEM-Scores außerhalb der KS-Stelle miteinbeziehen. Eine genaue Erklärung für die Berechnung des Mittelwerts der Differenz des TERbase-Scores und hLEPOR-Scores sowie seine Bedeutung ist unter §4.5.6.4 „Basis des Vergleichs vor-KS vs. nach-KS zur Ermittlung des KS-Einflusses“ aufgeführt.

[73] Näheres zu den AEMs TERbase und hLEPOR unter §4.5.6.3 „Auswahl der automatischen Evaluationsmetriken“.

5.5.10 Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

Unter diesem Analysefaktor wurde untersucht, inwiefern die Differenz der Scores von TERbase und hLEPOR mit der Differenz in der allgemeinen Qualität korreliert (Abbildung 5.157).



Differenz = Szenario nach KS *minus* Szenario vor KS

Abbildung 5.157: Differenz der allgemeinen Qualität und Differenz der AEM-Scores auf MÜ-Systemebene

Bei allen MÜ-Systemen erwies die Spearman-Korrelationsanalyse hochsignifikante mittlere und starke Korrelationen zwischen den Differenzen der Scores von TERbase und hLEPOR und der Differenz der allgemeinen Qualität[74] (Tabelle 5.178).

Somit bestätigen sich die Ergebnisse der Humanevaluation und die der automatischen Evaluation gegenseitig, denn eine starke positive Korrelation besagt, dass ein positiver Effekt der KS-Anwendung auf Systemebene sich sowohl durch einen erhöhten Score in der Humanevaluation als auch einen verbesserten AEM-Score in der automatischen Evaluation zeigte. Umgekehrt gingen auch bei einem negativen Effekt der KS-Anwendung die Scores beider Evaluationen zurück.

[74]Die allgemeine Qualität ist der Mittelwert der Stilqualität und der Inhaltsqualität, da bei der Ermittlung dieser Korrelation keine Unterscheidung zwischen den beiden Qualitätsattributen notwendig ist.

Tabelle 5.178: Korrelation zwischen den Differenzen der AEM-Scores und den Qualitätsdifferenzen

	Korrelation zw. Differenz allg. Qualität und Differenz des ***TERbase-Scores*** (nach KS *minus* vor KS)			Korrelation zw. Differenz allg. Qualität und Differenz des ***hLEPOR-Scores*** (nach KS *minus* vor KS)		
	N	**p**	**ρ**	**N**	**p**	**ρ**
HMÜ Bing	139	< ,001	,589	139	< ,001	,580
GNMÜ	194	< ,001	,606	194	< ,001	,611
RBMÜ Lucy	156	< ,001	,491	156	< ,001	,495
SMÜ SDL	153	< ,001	,418	153	< ,001	,436
HMÜ Systran	133	< ,001	,518	133	< ,001	,476

p: Signifikanz ρ: Korrelationskoeffizient
schwache Korrelation (ρ >=0,1) mittlere Korrelation (ρ >= 0,3) starke Korrelation (ρ >= 0,5)

Außerdem visualisiert Abbildung 5.157 noch einmal deutlich, wie gering die Qualitäts- und AEM-Differenzen vor vs. nach KS bei Google Translate und Systran im Vergleich zu den anderen Systemen ausfielen: bei Google Translate aufgrund der kleinen Veränderung in den vielen fehlerfreien MÜ bzw. bei Systran aufgrund der kleinen Veränderung in den vielen fehlerhaften MÜ in den beiden Szenarien.

5.5.11 Analyse auf MÜ-Systemebene: Validierung der Hypothesen

Um die vorgestellten Ergebnisse der Systeme auf die Forschungsfragen der Studie zurückzuführen, listet dieser Abschnitt die zugrunde liegenden Hypothesen der Forschungsfragen zusammen mit einer Zusammenfassung der obigen Ergebnisse in tabellarischer Form auf. Für einen schnelleren Überblick steht (+) für eine Verbesserung bzw. einen Anstieg z. B. im Sinne eines Qualitätsanstiegs, verbesserter AEM-Scores oder eines Anstiegs der Fehleranzahl; (–) steht für einen Rückgang; die grüne Farbe symbolisiert eine signifikante Veränderung; <<>>

steht für eine starke Korrelation und <> für eine mittlere Korrelation.[75]

Vergleich der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Fehleranzahl nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Unterschied in der Fehleranzahl vor vs. nach der Anwendung der KS-Regeln.RBMÜ

Resultat: Nur für Bing und SDL wurde H0 abgelehnt und somit H1 bestätigt:

HMÜ Bing (–)	**SMÜ SDL (–)**	
RBMÜ Lucy (–)	**GNMÜ +**	**HMÜ Systran +**

Vergleich der Fehlertypen vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Kommen bestimmte Fehlertypen *bei einem bestimmten MÜ-System* vor bzw. nach der Anwendung der KS-Regeln vor?

H0 – Es gibt keinen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Unterschied in der Häufigkeit der einzelnen Fehlertypen vor vs. nach der Anwendung der KS-Regeln.

Resultat: H0 wurde abgelehnt und somit H1 bestätigt, und zwar nur für die hier aufgeführten Fehlertypen bei dem jeweiligen System:

HMÜ Bing:	**GNMÜ**	**RBMÜ Lucy:**	**SMÜ SDL:**	**HMÜ Systran:**
+ OR.1		**– OR.2**	**+ OR.1**	**+ OR.1**
– OR.2		**– SM.13**	**– OR.2**	**– LX.3**
– LX.3			**– LX.3**	**+ LX.4**
– GR.8			**– LX.4**	**– SM.13**
– GR.10			**– GR.9**	
			– GR.10	

[75] Schwache Korrelationen werden in dieser Übersicht nicht angezeigt.

Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Stil- und Inhaltsqualität der MÜ der KS-Stelle nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regeln.

Resultat: H0 wurde abgelehnt und H1 bestätigt, und zwar wie folgt: SQ nur in Bing, Google und SDL; CQ nur in Bing und SDL.

	Bing:	**Google:**	**Lucy:**	**SDL**	**Systran:**
SQ	+	−	+	+	+
CQ	+	−	+	+	+

Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln auf Annotationsgruppenebene

Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in der Stil- und Inhaltsqualität bei den einzelnen Annotationsgruppen nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Bei den Annotationsgruppen gibt es keinen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

H1 – Bei den Annotationsgruppen gibt es einen Qualitätsunterschied vor vs. nach der Anwendung der KS-Regel.

Resultat: Nur für die Gruppen RF und FR bei allen Systemen mit Ausnahme von zwei Fällen bei Google (CQ in RF und SQ in FR) wurde H0 abgelehnt und somit H1 bestätigt:

System \ AnnoGr.	FF		RF		FR		RR	
	SQ	CQ	SQ	CQ	SQ	CQ	SQ	CQ
HMÜ Bing	+	+	(–)	(–)	+	+	(–)	+
GNMÜ	(–)	(–)	(–)	(–)	+	+	(–)	(–)
RBMÜ Lucy	+	(–)	(–)	(–)	+	+	(–)	+
SMÜ SDL	+	+	(–)	(–)	+	+	(–)	(–)
HMÜ Systran	(–)	(–)	(–)	(–)	+	+	+	+

Vergleich der Qualität vor vs. nach der Anwendung der KS-Regeln

Fragestellung: Besteht *bei einem bestimmten MÜ-System* ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps (Fehleranzahl nach KS *minus* Fehleranzahl vor KS) und der Differenz der Stil- bzw. Inhaltsqualität (Qualität nach KS *minus* Qualität vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der Fehleranzahl eines bestimmten Fehlertyps und der Differenz der Stil- bzw. Inhaltsqualität.

Resultat: H0 wurde abgelehnt und somit H1 bestätigt, und zwar nur für den Zusammenhang zwischen den Qualitätswerten und den hier aufgeführten Fehlertypen bei dem jeweiligen System. Alle aufgelisteten Korrelationen sind negativ:

HMÜ Bing	GNMÜ	RBMÜ Lucy	SMÜ SDL	HMÜ Systran
OR.1<>SQ	GR.10<>SQ	OR.2<>SQ	LX.3<>SQ	SM.13<>SQ
GR.8<>SQ			LX.4<>SQ	
GR.10<>SQ			GR.8<>SQ	
			GR.10<<>>SQ	
LX.3<>CQ	GR.10<<>>CQ	SM.12<<>>CQ	LX.3<<>>CQ	GR.10<>CQ
GR.8<>CQ			GR.8<>CQ	SM.11<>CQ
GR.10<>CQ			GR.10<>CQ	SM.12<>CQ
			SM.12<>CQ	

Vergleich der AEM-Scores vor vs. nach der Anwendung der KS-Regel

Fragestellung: Gibt es *bei einem bestimmten MÜ-System* einen Unterschied in den AEM-Scores von TERbase bzw. hLEPOR nach der Anwendung der KS-Regeln im Vergleich zu vor der Anwendung?

H0 – Es gibt keinen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

H1 – Es gibt einen Unterschied in den AEM-Scores vor vs. nach der Anwendung der KS-Regeln.

Resultat: H0 wurde nicht abgelehnt und somit konnte H1 nicht bestätigt werden. Die Ergebnisse waren wie folgt:

	Bing	**Google**	**Lucy**	**SDL**	**Systran**
TERbase	(–)	(–)	(–)	+	(–)
hLEPOR	+	(–)	(–)	+	(–)

Korrelation zwischen den AEM-Scores-Differenzen und der Qualitätsdifferenz

Fragestellung: Besteht *bei einem bestimmten MÜ-System* ein Zusammenhang zwischen der Differenz der AEM-Scores in TERbase bzw. hLEPOR (Mittelwert der AEM-Scores nach KS *minus* Mittelwert der AEM-Scores vor KS) und der Differenz der allgemeinen Qualität (Qualität nach KS *minus* Qualität vor KS)?

H0 – Es besteht kein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

H1 – Es besteht ein Zusammenhang zwischen der Differenz der AEM-Scores und der Differenz der allgemeinen Qualität.

Resultat: H0 wurde abgelehnt und H1 bestätigt, und zwar für den Zusammenhang zwischen der allgemeinen Qualität und den AEM-Scores wie folgt. Alle aufgelisteten Korrelationen sind positiv:

	SMÜ Bing	**GNMÜ**	**RBMÜ Lucy**	**SMÜ SDL**	**HMÜ Systran**
TERbase	<<>> Q	<<>> Q	<> Q	<> Q	<<>> Q
hLEPOR	<<>> Q	<<>> Q	<> Q	<> Q	<> Q

5.5.12 Übersicht der Ergebnisse auf MÜ-Systemebene

Tabelle 5.179 bietet eine Übersicht über die Ergebnisse auf Systemebene.

Der Vergleich der fünf unterschiedlichen MÜ-Systeme vor vs. nach der Anwendung der neun KS-Regeln deckte folgendes Ergebnis auf: Bei den Systemen der älteren Ansätze stimmten die Ergebnisse mit denen mehrerer bisheriger Studien zu den älteren MÜ-Ansätzen überein, nämlich dass die KS-Anwendung den MÜ-Output verbessert (vgl. Nyberg & Mitamura 1996; Bernth 1999; Bernth & Gdaniec 2001: 208; Drugan 2013: 98; Drewer & Ziegler 2014: 196; Wittkowsky 2017: 92). Dies ist der Fall mit Ausnahme eines der beiden analysierten HMÜ-Systeme (Systran), dessen Fehleranzahl sowohl vor-KS als auch nach-KS sehr hoch war und nach-KS insignifikant stieg. Das NMÜ-System hingegen war in der Lage, unter allen Systemen die Übersetzung mit der geringsten Fehleranzahl und den höchsten Qualitätswerten im Sinne der Verständlichkeit, der Genauigkeit und des Stils sowohl vor als auch nach der Anwendung der KS-Regeln zu produzieren. Ferner führte die Anwendung der analysierten Regeln bei dem NMÜ-System nicht zu einer Qualitätsverbesserung. Im Gegenteil fielen vor der KS-Anwendung die Fehleranzahl niedriger und die Verständlichkeit-, Genauigkeits- und Stilbewertungen höher als nach der KS-Anwendung – mit einem statistisch signifikanten Wert im Falle der Stilbewertungen – aus. Wie bisherige Studien zeigen (vgl. Toral & Sanchez-Cartagena 2017), zählt die Flüssigkeit bzw. der Stil zu den Stärken des NMÜ-Ansatzes. Obwohl die KS im Allgemeinen die Verständlichkeit und nicht den Stil im Fokus hat, bietet die NMÜ auf Basis der erzielten Ergebnisse nicht nur eine verbesserte Verständlichkeit und Genauigkeit, sondern schlägt sich auch positiv im Stil nieder.

Tabelle 5.179: Übersicht der Ergebnisse auf MÜ-Systemebene

	Fehler-	Fehler-	Qualität		Fehlertypen <> Qualität		AEM-Scores		AEM-Scores <> Qualität	
	anzahl	typen			Stilqualität	Inhaltsqualität	TERbase	hLEPOR	TERbase	hLEPOR
HMÜ Bing	(–)	+ OR.1 – OR.2 – LX.3 – GR.8 – GR.10	+ SQ	+ CQ	*neg* OR.1<>SQ *neg* GR.8<>SQ *neg* GR.10<>SQ	*neg* LX.3<>CQ *neg* GR.8<>CQ *neg* GR.10<>CQ	(–)	+	*pos* TERbase <<>> Q	*pos* hLEPOR <<>> Q
GNMÜ	+		– SQ	– CQ	*neg* GR.10<>SQ	*neg* GR.10<<>>CQ	(–)	(–)	*pos* TERbase <<>> Q	*pos* hLEPOR <<>> Q
RBMÜ Lucy	(–)	– OR.2 – SM.13	+ SQ	+ CQ	*neg* OR.2<>SQ	*neg* SM.12<<>>CQ	(–)	(–)	*pos* TERbase <> Q	*pos* hLEPOR <> Q
SMÜ SDL	(–)	+ OR.1 – OR.2 – LX.3 – LX.4 – GR.9 – GR.10	+ SQ	+ CQ	*neg* LX.3<>SQ *neg* LX.4<>SQ *neg* GR.8<>SQ *neg* GR.10<<>>SQ	*neg* LX.3<<>>CQ *neg* GR.8<>CQ *neg* GR.10<>CQ *neg* SM.12<>CQ	+	+	*pos* TERbase <> Q	*pos* hLEPOR <> Q
HMÜ Systran	+	+ OR.1 – LX.3 + LX.4 – SM.13	+ SQ	+ CQ	*neg* SM.13<>SQ	*neg* GR.10<>CQ *neg* SM.11<>CQ *neg* SM.12<>CQ	(–)	(–)	*pos* TERbase <<>> Q	*pos* hLEPOR <> Q

SQ: Stilqualität CQ: Inhaltsqualität Q: allg. Qualität Signifikant (p < 0,5) Blank: nicht signifikant

<> mittlere Korrelation (ρ >= 0,3) <<>> starke Korrelation (ρ >= 0,5) *neg*: negative Korrelation *pos*: positive Korrelation

6 Zusammenfassung und Diskussion der Ergebnisse

6.1 Einleitung

In diesem Kapitel werden die Studienergebnisse zusammenfassend diskutiert. Zunächst wird die festgestellte allgemeine positive Auswirkung der analysierten KS-Regeln erörtert. Im Anschluss wird die Auswirkung differenziert auf Regelebene (systemübergreifend) diskutiert. Auf einer tieferen Ebene wird der Effekt der einzelnen Regeln auf Systemebene (Analyse auf Regel- und Systemebene) zusammenfassend erörtert. Zum Schluss wenden wir uns dem Vergleich der vier MÜ-Ansätze (RBMÜ, SMÜ, HMÜ und NMÜ) jeweils am Beispiel des verwendeten Systems nach der Anwendung aller analysierten Regeln (d. h. regelübergreifend) zu.

6.2 Allgemeine Auswirkung der KS-Regeln

Die allgemeine Auswirkung der KS-Regeln auf den MÜ-Output stimmt mit den Ergebnissen früherer empirischer und theoretischer Studien (vgl. Nyberg & Mitamura 1996; Bernth 1999; Bernth & Gdaniec 2001: 208; Drugan 2013: 98; Drewer & Ziegler 2014: 196; Wittkowsky 2017: 92) überein, bei denen festgestellt wurde, dass die KS-Anwendung den MÜ-Output aus unterschiedlichen Perspektiven verbessert. Die Ergebnisse aller angewandten Methoden bestätigten systemübergreifend die kollektive positive Wirkung aller analysierten Regeln auf den MÜ-Output:

In der Fehlerannotation sank die Fehleranzahl nach der Anwendung der Regeln signifikant. Sechs Fehlertypen nahmen nach der Anwendung der KS-Regeln in der folgenden Reihenfolge signifikant ab: (1) SM.13 „Kollokationsfehler“, (2) LX.3 „Wort ausgelassen“, (3) GR.8 „Falsches Verb“, (4) OR.2 „Großschreibung“, (5) GR.10 „Falsche Wortstellung“ und (6) GR.9 „Kongruenzfehler“. Vergleichbar mit dem Ergebnis einer vorherigen Studie (vgl. Kirchhoff u. a. 2014) war der Wortstellungsfehler (GR.10) der einzige Fehlertyp, dessen Rückgang mit dem Anstieg

beider Qualitätswerte (SQ und CQ) signifikant korrelierte, obwohl der Rückgang im Fall des Wortstellungsfehlers (GR.10) nicht der höchste unter den Fehlertypen war. Im Übrigen bestand nur eine weitere Korrelation zwischen dem Rückgang des Auslassungsfehlers (LX.3) und dem Anstieg der CQ. Gleichzeitig nahmen zwei Fehlertypen signifikant zu (LX.6 „Konsistenzfehler“ und OR.1 „Zeichensetzung“), jedoch bestand keine Korrelation zwischen ihrem Anstieg und der Qualitätsveränderung.

In der Humanevaluation stiegen sowohl die Stil- als auch die Inhaltsqualität signifikant, wobei der Anstieg der Inhaltsqualität (Verständlichkeit und Genauigkeit) höher war. In der automatischen Evaluation nahmen beide AEM-Scores leicht zu. Außerdem zeigte der Spearman-Korrelationstest eine signifikante starke positive Korrelation zwischen der Differenz (nach-KS *minus* vor-KS) der allgemeinen Qualität und der der AEM-Scores, was darauf hindeutet, dass der Anstieg der allgemeinen Qualität mit der Verbesserung der AEM-Scores einherging.

Die unterschiedlichen Anstiegsniveaus der Stil- und Inhaltsqualität nach der Anwendung der KS-Regeln werfen jedoch die Frage auf, in welchen Fällen sich eine deutliche Steigerung der Inhaltsqualität gegenüber der Stilqualität zeigte. Diese Frage lässt sich auf Regelebene beantworten.

6.3 Systemübergreifende Auswirkung der KS auf Regelebene

Für einen Überblick über die Ergebnisse auf Regelebene betrachten wir im Folgenden die einzelnen analysierten Regeln aus drei Perspektiven, nämlich unter Berücksichtigung der Fehleranzahlveränderungen, der Annotationsgruppen sowie der Qualitätsveränderungen.

6.3.1 Überblick über die Fehleranzahlveränderungen der einzelnen Regeln

Die Fehleranalyse ergab, dass die Fehleranzahl nach der Anwendung von sieben der analysierten neun Regeln sank, wobei dieser Fehleranzahlrückgang nur bei fünf von diesen Regeln signifikant war. Diese Regeln sind absteigend nach der Prozentzahl des Fehleranzahlrückgangs sortiert, wie folgt (Abbildung 6.1): „fvg – Funktionsverbgefüge vermeiden“ mit einem Rückgang von 56,7 %; „kos – Konditionalsätze mit ‚Wenn‘ einleiten“ mit einem Rückgang von 56,5 %; „anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ mit einem

Rückgang von 46,0 %; „per – Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden" mit einem Rückgang von 37,2 % und „prä – Überflüssige Präfixe vermeiden" mit einem Rückgang von 35,6 %.

Bei der achten „pak – Partizipialkonstruktionen vermeiden" und neunten Regel „pas – Passiv vermeiden" stieg die Fehleranzahl nach der Regelanwendung (Abbildung 6.1), wobei der Anstieg nur bei der Regel „pak – Partizipialkonstruktionen vermeiden" (26,4 %) signifikant war.

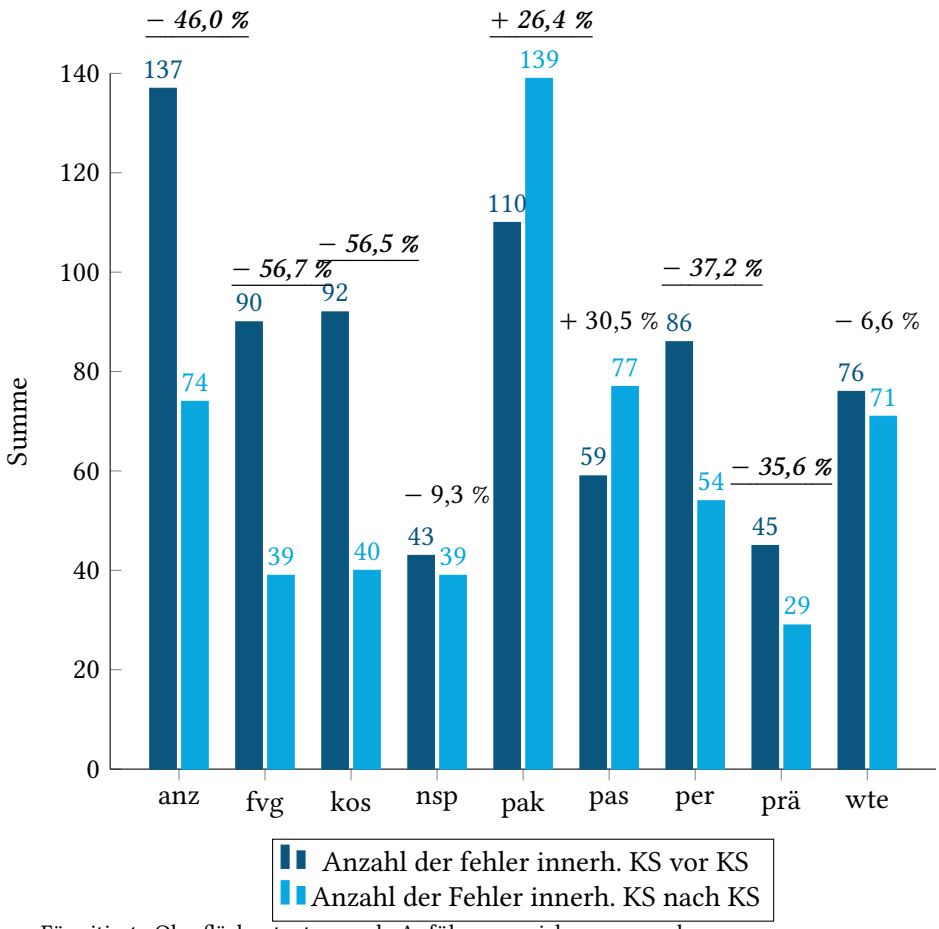
Eine genauere Aussage, ob diese Fehleranzahlrückgänge bzw. Fehleranzahlanstiege mit einer Qualitätsverbesserung bzw. Qualitätsverschlechterung einhergingen, konnte an dieser Stelle nicht getroffen werden. Dies bedarf einer tieferen Analyse, in der die Ergebnisse der Fehleranalyse mit denen der Humanevaluation (Tabelle 6.1) sowie die Ergebnisse der Humanevaluation mit denen der automatischen Evaluation (Abbildung 6.4) trianguliert werden.

6.3.2 Überblick über die Annotationsgruppen der einzelnen Regeln

In der Analyse auf Annotationsgruppenebene wurden die Ergebnisse nach dem Vorhandensein bzw. Nichtvorhandensein von MÜ-Fehlern in vier Gruppen aufgeteilt: RR, FF, RF und FR.[1] Nach dieser Analyse zeigten die Regeln eine ziemlich begrenzte positive Wirkung, nämlich ausschließlich bei der FR-Gruppe (d. h. MÜ beinhaltet vor der Regelanwendung Fehler und ist nach der Regelanwendung fehlerfrei). Diese Gruppe liegt lediglich zwischen 8 % (Regel „Passiv vermeiden") und 31 % (Regel „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden"), Abbildung 6.2.

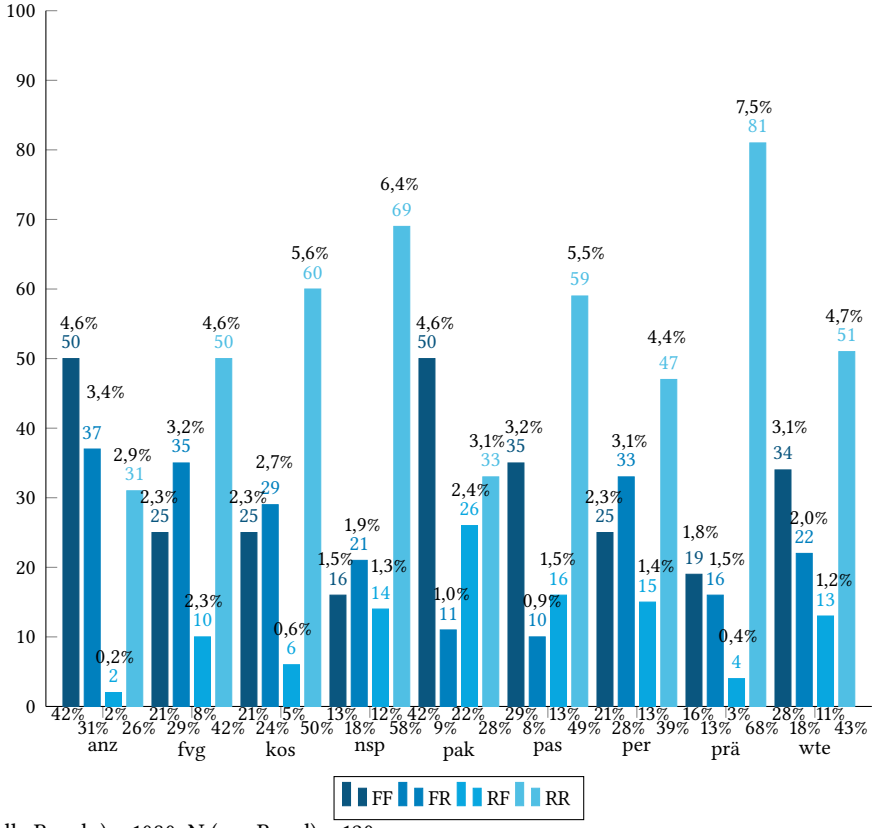
In der RF-Annotationsgruppe ist die KS-Auswirkung eindeutig negativ. Die bei allen Regeln dominierenden Annotationsgruppen waren RR und FF. Da die Übersetzungen sowohl vor als auch nach der Anwendung der jeweiligen Regel fehlerfrei (RR-Gruppe) oder fehlerhaft (FF-Gruppe) waren, kann eine positive Auswirkung einer Regel nur dann gerechtfertigt sein, wenn die Qualitätswerte dieser beiden Gruppen nach der Regelanwendung stiegen. Eine Qualitätssteigerung in der RR-Gruppe würde bedeuten, dass die Qualität einer fehlerfreien MÜ nach KS höher als die einer fehlerfreien MÜ vor KS sei (z. B. durch eine stilistische Verbesserung). Ebenso würde eine Qualitätssteigerung in der FF-Gruppe bedeuten, dass beim Vergleich zweier fehlerhaften Übersetzungen vor und nach

[1] Die Daten wurden binär bzw. dichotom aufgeteilt (keine Fehler aufgetreten ‚0'; Fehler aufgetreten ‚1'), daraus wurden vier Annotationsgruppen in Bezug auf die *KS-Stelle* gebildet: (1) RR: MÜ ist vor und nach der Anwendung der KS-Regel fehlerfrei; (2) FF: MÜ beinhaltet vor und nach der Anwendung der KS-Regel Fehler; (3) RF: MÜ ist nur vor der Anwendung der KS-Regel fehlerfrei; (4) FR: MÜ ist nur nach der Anwendung der KS-Regel fehlerfrei.



anz: Für zitierte Oberflächentexte gerade Anführungszeichen verwenden
fvg: Funktionsverbgefüge vermeiden
kos: Konditionalsätze mit ‚Wenn' einleiten
nsp: Eindeutige pronominale Bezüge verwenden
pak: Partizipial-konstruktionen vermeiden
pas: Passiv vermeiden
per: Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden
prä: Überflüssige Präfixe vermeiden
wte: Keine Wortteile weglassen
Signifikante Differenz vor vs. nach KS

Abbildung 6.1: Summe der Fehleranzahl vor vs. nach KS auf Regelebene



N (alle Regeln) = 1080; N (pro Regel) = 120
Die oben angezeigten Prozentzahlen sind auf Basis des Gesamtdatensatzes aller analysierten Regeln (N = 1080) berechnet.
Die untenstehenden Prozentzahlen sind auf Regelebene (N = 120) berechnet.

Abbildung 6.2: Aufteilung der Annotationsgruppen auf Regelebene

KS die Qualität der fehlerhaften MÜ nach KS höher wäre (z. B. aufgrund des Auftretens eines weniger schwerwiegenden Fehlertyps oder des Rückgangs der Fehleranzahl).

Die Triangulation der Ergebnisse der Fehlerannotation und der Humanevaluation deckte auf, wie sich jede Regel auf Annotationsgruppenebene auf die Stil- und Inhaltsqualität auswirkte (Tabelle 6.1). In der Tabelle wird im Wesentlichen der Mittelwert (M) der Qualitätsveränderung (Diff. SQ, Diff. CQ und Diff. Q berechnet als nach-KS *minus* vor-KS) bei jeder Annotationsgruppe sowie die Signifikanz (p) dieser Veränderung nach dem Wilcoxon-Test aufgeführt. Der grüne Hintergrund verdeutlicht die signifikanten Fälle.

Nur zwei Regeln wirkten sich positiv auf die MÜ-Qualität aus. Das lässt sich in Tabelle 6.1 – wie oben erwähnt – anhand der signifikant positiven Qualitätsdifferenzmittelwerte (positiver M bei $p < 0{,}05$) bei den beiden dominanten Gruppen RR und FF sowie der Gruppe FR erkennen. Eine signifikant positive Qualitätsdifferenz bei der Gruppe RR deutet darauf hin, dass bei zwei fehlerfreien MÜ vor und nach der KS-Anwendung die MÜ-Qualität nach KS höher bewertet wurde (z. B. durch eine stilistische Verbesserung). Ebenfalls zeigt ein signifikant positiver Qualitätswert bei der Gruppe FF, dass bei zwei fehlerhaften MÜ vor und nach der KS-Anwendung die MÜ-Qualität nach KS höher bewertet wurde (z. B. aufgrund des Rückgangs der Fehleranzahl oder des Auftretens von vergleichbar weniger gravierenden Fehlern). Diese zwei Regeln sind:

Erstens – die Regel *„anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden"* war die einzige Regel, bei der die SQ- und CQ-Werte der FF- und RR-Gruppen nach der Implementierung der Regel signifikant anstiegen. Außerdem war die FR-Gruppe mit dem höchsten Prozentsatz (31 %) und die RF-Gruppe mit dem niedrigsten (2 %) vertreten. Orthografisch haben die Anführungszeichen die Aufgabe der „Kennzeichnung der Grenzen eines Einschubs innerhalb des Satzverbandes" (Nerius 2007: 253). Die Kennzeichnung der Oberflächentexte mithilfe der Anführungszeichen unterstützte demnach systemübergreifend dabei, die Oberflächentexte als spezifische Begriffe bzw. Mehrwortentitäten zu identifizieren und somit den Satz besser zu parsen. Dementsprechend hatte diese Regel einen eindeutigen positiven Einfluss auf die MÜ-Qualität.

Zweitens – die Regel *„per – Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden"*, bei der die SQ sich in der FF- und RR-Gruppe signifikant verbesserte, während die CQ keine signifikante Veränderung zeigte. Darüber hinaus nahmen sowohl die SQ als auch die CQ in der FR-Gruppe signifikant zu. Systemübergreifend konnte die Passiversatzkonstruktion (vor-KS) in 52 % der Fälle (Gruppe RR und RF) fehlerfrei übersetzt werden. Die Konstruktion „sein + zu + Infinitiv" ist im Deutschen eine Form des Passiversatzes für ein Passiv mit Modalverb, die ins

Tabelle 6.1: Differenz der Stil- und Inhaltsqualität nach der Anwendung jeder KS-Regel auf Annotationsgruppenebene

		FF			FR			RF			RR		
		Diff. SQ	Diff. CQ	Diff. Q	Diff. SQ	Diff. CQ	Diff. Q	Diff. SQ	Diff. CQ	Diff. Q	Diff. SQ	Diff. CQ	Diff. Q
anz	M	0,47	0,27	0,37	0,69	0,84	0,77	− 0,38	0	− 0,19	0,32	0,05	0,18
	p	< 0,001	0,003	< 0,001	< 0,001	0,001	< 0,001	1	1	1	0,001	*0,273*	0,005
	N	30	30	30	24	24	24				19	19	19
	z	− 3,711	− 2,998	− 3,659	− 3,950	− 3,402	− 3,787				− 3,282	− 1,096	− 2,825
fvg	M	0,14	0,14	0,14	0,85	0,75	0,80	− 0,16	− 0,50	− 0,33	0,15	0,03	0,09
	p	*0,420*	*0,636*	*0,421*	< 0,001	< 0,001	< 0,001	*0,416*	*0,107*	*0,058*	*0,140*	*0,926*	*0,190*
	N	17	17	17	25	25	25	7	7	7	35	35	35
	z	− 0,807	− 0,474	− 0,805	− 4,205	− 3,884	− 4,043	− 0,813	− 1,612	− 1,892	− 1,477	− 0,093	− 1,312
kos	M	0,30	0,38	0,34	0,71	1,28	0,99	− 0,71	− 1,04	− 0,88	− 0,10	0,01	− 0,04
	p	*0,090*	*0,278*	*0,157*	< 0,001	< 0,001	< 0,001	*0,285*	*0,285*	*0,285*	*0,620*	*0,795*	*0,229*
	N	14	14	14	23	23	23	3	3	3	44	44	44
	z	− 1,694	− 1,084	− 1,414	− 4,143	− 4,209	− 4,201	− 1,069	− 1,069	− 1,069	− 0,222	− 0,260	− 1,204
nsp	M	− 0,15	− 0,08	− 0,11	0,31	0,68	0,50	− 0,50	− 0,48	− 0,49	− 0,05	0,10	0,02
	p	*0,443*	*0,799*	*0,735*	0,011	0,002	0,002	0,012	0,012	0,012	*0,220*	*0,086*	*0,795*
	N	8	8	8	12	12	12	10	10	10	47	47	47
	z	− 0,768	− 0,254	− 0,338	− 2,547	− 3,072	− 3,065	− 2,509	− 2,313	− 2,527	− 1,227	− 1,718	- 0,260
pak	M	− 0,19	0,05	− 0,07	0,44	0,82	0,63	− 0,73	− 0,52	− 0,63	− 0,38	− 0,13	− 0,25
	p	0,008	*0,411*	*0,507*	*0,107*	0,021	*0,050*	< 0,001	0,012	< 0,001	< 0,001	0,020	< 0,001
	N	37	37	37	9	9	9	23	23	23	28	28	28
	z	− 2,638	− 0,823	− 0,664	− 1,612	− 2,314	− 1,958	− 4,114	− 2,526	− 4,021	− 3,540	− 2,322	− 3,618
pas	M	− 0,34	− 0,16	− 0,25	0,54	0,58	0,56	− 0,99	− 1,93	− 1,46	− 0,13	0,05	− 0,04
	p	0,010	*0,177*	0,019	*0,114*	*0,080*	0,075	0,003	0,002	0,002	*0,061*	*0,427*	*0,383*
	N	25	25	25	6	6	6	12	12	12	40	40	40
	z	− 2,566	− 1,351	− 2,345	− 1,581	− 1,753	− 1,782	− 2,940	− 3,062	− 3,061	− 1,876	− 0,794	− 0,872
per	M	0,33	0,27	0,30	1,29	1,59	1,44	− 0,33	− 0,47	− 0,40	0,24	0,00	0,12
	p	0,003	*0,136*	0,012	< 0,001	< 0,001	< 0,001	0,016	0,007	0,001	< 0,001	*0,522*	0,002
	N	20	20	20	27	27	27	13	13	13	37	37	37
	z	− 3,021	− 1,492	− 2,505	− 4,463	− 4,544	− 4,542	− 2,417	− 2,680	− 3,192	− 3,737	− 0,640	− 3,167
prä	M	0,10	0,08	0,09	0,65	0,78	0,71	− 0,31	− 0,88	− 0,59	− 0,05	− 0,03	− 0,04
	p	*0,109*	*0,765*	*0,439*	0,002	0,001	0,001	*0,180*	*0,180*	*0,180*	*0,148*	*0,828*	*0,122*
	N	11	11	11	17	17	17	2	2	2	62	62	62
	z	− 1,602	− 0,299	− 0,775	− 3,133	− 3,462	− 3,410	− 1,342	− 1,342	− 1,342	− 1,448	− 0,217	− 1,548
wte	M	− 0,27	− 0,29	− 0,28	0,20	0,52	0,36	− 0,50	− 1,05	− 0,78	− 0,35	− 0,06	− 0,21
	p	0,012	*0,177*	0,018	*0,266*	0,001	0,010	0,028	0,004	0,004	< 0,001	*0,078*	< 0,001
	N	24	24	24	16	16	16	12	12	12	35	35	35
	z	− 2,521	− 1,351	− 2,361	− 1,113	− 3,221	− 2,582	− 2,201	− 2,875	− 2,904	− 4,575	− 1,765	− 4,175

SQ: Stilqualität; CQ: Inhaltsqualität; Q: Allgemeine Qualität (Mittelwert der SQ und CQ); Diff. SQ = SQ vor − SQ nach KS; analog dazu sind Diff. CQ und Diff. Q.
insignifikante Werte p ≥ 0,05; signifikante Werte p < 0,05

Englische als reguläres Passiv übersetzt wird (vgl. Teich 2003: 92; König & Gast 2012: 161). Eine fehlerfreie Übersetzung des Imperativs (nach-KS) hingegen war in 67 % der Fälle (Gruppe RR und FR) unproblematisch. Durch die ähnlichen Prozentsätze der richtigen Fälle in beiden Szenarien (52 % vor-KS vs. 67 % nach-KS) wurde die CQ nicht signifikant beeinflusst. Dennoch konnte die Verwendung des Imperativs (nach-KS) die MÜ überwiegend auf stilistischer Ebene verbessern. Ferner akzentuiert die zu 48 % fehlerhafte MÜ vor der Regelanwendung (Gruppe FR und FF) die Schwierigkeit der MÜ dieser Passiversatzkonstruktion, die auf den kontrastiven Unterschied zwischen dem Deutschen und Englischen (vgl. Teich 2003: 93) zurückgeführt werden kann.

Negative Auswirkungen von KS-Regeln konnten bei drei Regeln beobachtet werden. Das lässt sich in Tabelle 6.1 anhand der signifikant negativen Qualitätsdifferenzmittelwerte (negativer M bei $p < 0{,}05$) bei den beiden dominanten Gruppen RR und FF erkennen. Eine signifikant negative Qualitätsdifferenz bei der Gruppe RR weist darauf hin, dass bei zwei fehlerfreien MÜ vor und nach der KS-Anwendung die MÜ-Qualität nach KS niedriger bewertet wurde (z. B. aufgrund eines schlechteren Stils). Ebenfalls zeigt eine signifikant negative Qualitätsdifferenz bei der Gruppe FF, dass bei zwei fehlerhaften MÜ vor und nach der KS-Anwendung die MÜ-Qualität nach KS niedriger bewertet wurde (z. B. aufgrund der Zunahme der Fehleranzahl oder des Auftretens von vergleichbar gravierenden Fehlern). Diese drei Regeln sind:

Erstens – in der Regel *„pak – Partizipialkonstruktion vermeiden"* war die RF-Gruppe (22 %) mehr als doppelt so hoch vertreten wie FR (9 %). Die Untersuchung der FF- (42 %) und RR-Gruppen (28 %) zeigt, dass die SQ in FF signifikant fiel und sowohl die SQ als auch die CQ in RR signifikant abnahmen (d. h. bei einem Vergleich zweier fehlerfreien MÜ vor und nach der Regelanwendung sind die SQ und CQ der Partizipialkonstruktionen (vor-KS) höher). In der FR-Gruppe stieg nur die CQ signifikant an, während der SQ-Anstieg marginal war. Die Ergebnisse weisen auf die Komplexität der MÜ von Partizipialkonstruktionen hin (ca. die Hälfte der Fälle vor-KS beinhalteten Fehler), denn sie erschweren den Systemen das Parsen (vgl. Reuther 2003). Gleichzeitig zeigen die Ergebnisse, dass die Ersetzung einer Partizipialkonstruktion durch einen Nebensatz mit einer überwiegend stilistischen Qualitätsverschlechterung sowie einer verringerten Verständlichkeit der MÜ verbunden war. Somit konnte die Regel die maschinelle Übersetzbarkeit nicht fördern, wie es von Bernth & Gdaniec (2001) erwartet wird.

Zweitens – bei der Regel *„pas – Passiv vermeiden"* fiel die RF-Gruppe (13 %) größer als die FR (8 %) aus. Im Gegensatz zur Regel „pak" war die RR-Gruppe (49 %) jedoch viel höher vertreten als FF (29 %), was zeigt, dass die Systeme in beiden Szenarien (Passiv und Aktiv) fast die Hälfte der Sätze korrekt übersetzen

konnten. In der FF-Gruppe führte die Verwendung des Aktivs (nach-KS) zu einer signifikant niedrigeren SQ sowie einem insignifikanten Rückgang der CQ. Selbst in der FR-Gruppe waren die beobachteten Anstiege bei SQ und CQ nicht signifikant. In der RR-Gruppe änderten sich die Qualitätswerte bei der Verwendung des Aktivs im Vergleich zum Passiv leicht (ein insignifikanter Rückgang der SQ und eine insignifikante Zunahme der CQ). Eine genaue Betrachtung der Aufteilung der Annotationsgruppen und der Veränderung der Qualitätswerte zeigt, dass die Systeme in der Lage waren, mehr als die Hälfte beider Varianten fehlerfrei zu übersetzen. Zwei Faktoren spielen eine wesentliche Rolle bei dem beobachteten Einfluss der Regel auf die Qualität: (1) welche der beiden Varianten aus Sicht der Bewerter auf stilistischer Ebene für die Handlung erforderlich war (d. h. inwiefern es erforderlich war, den Leser direkt anzusprechen (Aktiv) bzw. die Handlung in den Vordergrund zu rücken (Passiv)); und (2) ob der Bewerter eher das Aktiv oder das Passiv gewohnt ist (vgl. Baumert & Verhein-Jarren 2012: 68). Nach Baumert & Verhein-Jarren (2012: 68f.) sind die Leser der unterschiedlichen technischen Dokumentationen die eine oder andere Formulierung gewohnt, z. B. erwarten lesende Ingenieure Passivsätze; Leser von Zeitschriftenbeiträgen und Präsentationen hingegen das Aktiv.

Drittens – Regel *„wte – Keine Wortteile weglassen“*: Auf orthografischer Ebene gibt es keine Regel zur Verwendung bzw. Nicht-Verwendung des Ergänzungsstriches. Entscheidet der Schreibende sich dafür, keine Wortteile wegzulassen, kann dies „zwar gegen stilistische Normen verstoßen, stellt aber keinen orthographischen Fehler dar“ (Nerius 2007: 191). Bei dieser Regel stieg in der FR-Gruppe nur die CQ signifikant an. Die Regelanwendung hatte einen signifikanten negativen Einfluss auf die SQ, der auch in den Gruppen RF, FF und RR festgestellt wurde. Dies zeigte, dass die Wiederholung, die durch die Verwendung vollständiger Wörter anstelle ihrer reduzierten Formen entsteht (z. B. *Start- und Endpunkt* → *Startpunkt und Endpunkt*), die Verständlichkeit erhöhen konnte und gleichzeitig stilistisch nicht akzeptabel war. Dementsprechend kann der technische Redakteur je nach Textsorte entscheiden, welcher der beiden Aspekte (Verständlichkeit vs. Stil) priorisiert werden soll.

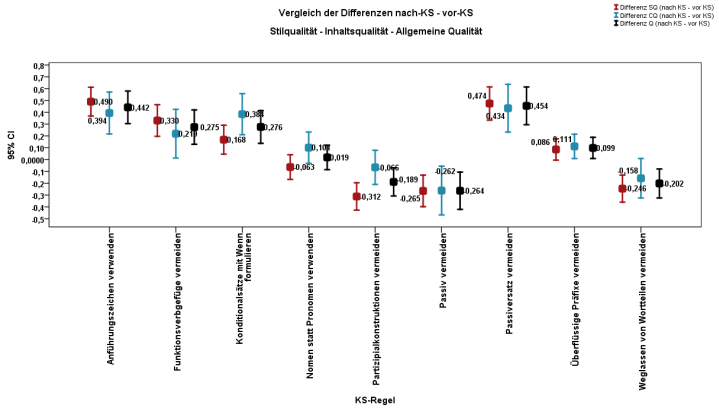
Die übrigen Regeln („fvg – Funktionsverbgefüge vermeiden“, „kos – Konditionalsätze mit ‚Wenn‘ einleiten“, „nsp – Eindeutige pronominale Bezüge verwenden“ und „prä – Überflüssige Präfixe vermeiden“) konnten auf dieser Analyseebene keinen signifikanten Einfluss auf die MÜ-Qualität zeigen. Alle Qualitätswerte dieser vier Regeln in den FF- und RR-Gruppen waren nicht signifikant.

Somit ergab die Analyse der triangulierten Daten auf Annotationsgruppenebene folgendes Ergebnis: Anders als die Ergebnisse der allgemeinen Auswirkung der KS-Regeln, die sich als positiv erwies (siehe §6.2), zeigen nur zwei der neun

Regeln eine positive Auswirkung auf den MÜ-Output. Eine weitere Analyse der Qualitätsveränderungen war dementsprechend erforderlich, um sicherzustellen, ob nur zwei Regeln die Qualität verbessern konnten.

6.3.3 Überblick über die Qualitätsveränderungen bei den einzelnen Regeln

Die Analyse der Qualitätsveränderungen basierend auf den Human- und automatischen Evaluationen bestätigt weitgehend die Ergebnisse, die auf Annotationsgruppenebene (Tabelle 6.1) erzielt wurden. Abbildung 6.3 liefert einen Überblick über die Qualitätsveränderungen auf Regelebene:



Qualitätsdifferenz = Qualitätswert nach KS *minus* Qualitätswert vor KS

Abbildung 6.3: Mittelwert der Qualitätsdifferenz auf Regelebene

Die Regeln „anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“ und „per – Konstruktionen mit ‚sein + zu + Infinitiv‘ vermeiden“ zeigten wiederholt einen signifikanten positiven Einfluss auf die MÜ-Qualität. Die Scores von TERbase und hLEPOR verbesserten sich. Gleichzeitig wurde basierend auf der Humanevaluation ein signifikanter positiver Einfluss auf die SQ und CQ auf Regelebene festgestellt:

Bei *„anz – Für zitierte Oberflächentexte*[2] *gerade Anführungszeichen verwenden“* war der positive Einfluss auf den Stil auf die klare orthografische Darstellung der Oberflächentexte zurückzuführen. Darüber hinaus wurde durch die Verwendung der Anführungszeichen die Eignung der Übersetzung für die Intention ih-

[2] Als Oberflächentexte gelten „alle Texte einer Softwareoberfläche oder Texte, die sich auf einem Gerät befinden“ tekom 2013: 117).

res Inhalts verbessert. Die festgestellten höheren Qualitätswerte bestätigen damit, dass eine korrekte Anwendung der Anführungszeichen eine korrekte Übersetzung fördern kann, wie es von der (tekom 2013: 118) beabsichtigt wird. Orthografisch haben die Anführungszeichen die Aufgabe, die Grenze einer fremden Äußerung innerhalb des Satzverbandes zu kennzeichnen (Nerius 2007: 253f.). Die Kennzeichnung der Oberflächentexte mithilfe der Anführungszeichen konnte daher die Systeme bei der Tokenisierung und dem Parsen unterstützen, was zur Verbesserung der MÜ beitrug. Gleichzeitig wurde die Aussage von McMurrey (2006), dass die Hervorhebung keine traditionelle Verwendung der Anführungszeichen im Englischen darstelle, in den Kommentaren der Bewerter thematisiert, auch wenn ihre vergebenen Qualität-Scores nach der Regelanwendung dadurch nicht negativ beeinflusst wurden. Eine Identifizierung der Oberflächentexte als feste bzw. spezifische Begriffe kann aber durch das Terminologiemanagement realisiert werden. Kunden- und produktspezifischen Begriffe werden in der Praxis als Termini eingeführt und entsprechend einer vom jeweiligen Unternehmen hinterlegten Terminologiedatenbank konsistent übersetzt (vgl. Volk 2018). Auf diese Weise würde die Identifizierung der Termini ebenfalls ein korrektes Parsen unterstützen und somit wäre die Regelanwendung zum Zwecke der maschinellen Übersetzbarkeit nicht mehr erforderlich (Näheres dazu unter §6.4).

In Bezug auf *„per – Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden"* fanden die Bewerter die Verwendung des Imperativs anstelle der Konstruktion „sein + zu + Infinitiv" stilistisch besser, da der Leser mit dem Imperativ direkt angesprochen und zum Handeln angeregt wurde. In Bezug auf die CQ wurde sowohl die Genauigkeit als auch die Klarheit nach der Regelanwendung erhöht, während der Effekt auf die Klarheit höher war. In dieser Hinsicht konnte – systemübergreifend – der gewünschte Effekt der Regelanwendung in der Ausgangssprache, auf den die tekom (2013: 86) und Congree (2018) abzielen, nämlich den Leser direkt anzusprechen und damit eine schnelle und richtige Handlungsumsetzung zu fördern, in der Zielsprache beobachtet werden.

In Bezug auf die Regel *„fvg – Funktionsverbgefüge vermeiden"*, deckte die Fehlerannotation, obwohl die Analyse der Annotationsgruppen keine wesentliche Qualitätssteigerung widerspiegelte (nur die Qualität in FR stieg signifikant an, siehe Tabelle 6.1), auf, dass die Verwendung des bedeutungstragenden Verbs anstelle des Funktionsverbgefüges zur semantischen und lexikalischen Verbesserung der MÜ beiträgt. Die ausdrucksschwachen Verben waren für die Systeme oft ambig. Insbesondere bei präpositionalen Funktionsverbgefügen (z. B. ‚zur Anwendung kommen'), Funktionsverbgefügen mit Komposita (z. B. ‚Fleckenbehandlung durchführen') und Funktionsverbgefügen ohne englisches Pendant

(z. B. ‚Einstellungen vornehmen') unterstützte die Regel die Systeme dabei, Transfer- und Parsing-Probleme zu bewältigen. So wurde die MÜ im Rahmen der Humanevaluation nach der Regelanwendung als geeigneter für die Satzintention und verständlicher bewertet. Sowohl bei der Humanevaluation als auch bei der automatischen Evaluation wurde nach der Regelanwendung eine signifikante Steigerung der MÜ-Qualität (SQ, CQ und beide AEM-Scores) festgestellt. Systemübergreifend stehen die Studienergebnisse in Einklang mit vorherigen Studien, die das Vermeiden des Funktionsverbgefüges bzw. ausdrucksschwacher Verben zur Reduzierung der Ambiguität, Vereinfachung der Satzstruktur und somit Verbesserung des Textverständnisses und der Übersetzung empfehlen (Siegel 2011; Congree 2018).

In Bezug auf *„kos – Konditionalsätze mit ‚Wenn' einleiten"* verbesserten sich SQ und CQ auf Annotationsgruppenebene nicht signifikant (siehe Tabelle 6.1). Auch die automatische Evaluation (sowohl TERbase als auch hLEPOR) zeigte nach der Regelanwendung nur eine geringe Qualitätssteigerung. Das liegt daran, dass 50 % der Sätze sowohl vor als auch nach der Regelanwendung fehlerfrei übersetzt wurden (Gruppe RR). Nach der Regelanwendung konnten die Fehler nur in 24 % der Sätze behoben werden (Gruppe FR). Diese Verbesserung war in den Ergebnissen der Humanevaluation im Sinne einer signifikant höheren SQ und CQ nachweisbar. Das Auslassen der Konjunktion ‚Wenn', das grammatisch im Deutschen aber nicht im Englischen möglich ist, führte zu Problemen bei der Syntaxanalyse. Die Regelanwendung ermöglichte eine bessere Syntaxanalyse (in Übereinstimmung mit der Feststellung der Komplexität der MÜ elliptischer Konstruktionen von Reuther 2003), daher wurden die Qualitätsbewertungen in Bezug auf Genauigkeit, Klarheit und Idiomatik erhöht.

Die Auswirkungen der Regeln „pak – Partizipialkonstruktion vermeiden", „wte – Keine Wortteile weglassen" und „pas – Passiv vermeiden" waren negativ:

Die Regel *„pak – Partizipialkonstruktion vermeiden"* wurde angewendet, indem basierend auf der Partizipialkonstruktion ein Nebensatz generiert wurde. Die Humanevaluation ergab, dass die MÜ der Partizipialkonstruktion idiomatischer, orthografisch korrekter und verständlicher als die des Nebensatzes war. Dementsprechend nahmen systemübergreifend sowohl die SQ als auch die CQ nach der Regelanwendung ab, wobei nur der SQ-Rückgang signifikant war. Die automatische Evaluation bestätigte dieses Ergebnis und zeigte eine signifikante Verschlechterung der Qualitätscores von TERbase und hLEPOR. Anders als von Bernth & Gdaniec (2001) erwartet, nämlich dass diese Regel die maschinelle Übersetzbarkeit fördern würde, konnte dies im Rahmen der Studie nicht bestätigt werden. Die Ergebnisse spiegeln wider, dass die Regel die Satzkomplexität

zwar reduzierte, die MÜ nach der Regelanwendung jedoch stilistisch kritisch und weniger verständlich ausfiel.

Im Fall von *„wte – Keine Wortteile weglassen“* empfanden die Evaluatoren die MÜ aufgrund der Substantivwiederholung (anstelle der reduzierten Form) als unnatürlich. So sank die SQ signifikant, während die CQ nicht signifikant abnahm. Darüber hinaus sanken die Bewertungen beider AEMs signifikant. Systemübergreifend zeigen die Ergebnisse somit, dass die MÜ von dieser Form der Ellipsen sich verbessert hat, was einen prägnanten und natürlichen Stil fördert. So kann das Unternehmen je nach Kritikalitätsgrad des Kontexts bei dem Terminologiemanagement individuell festlegen, inwiefern ein vollständiges Ausschreiben der Wörter zwecks der Verständlichkeit bzw. Eindeutigkeit erforderlich ist. Die Regel kann entsprechend über die Spanne zwischen der Prägnanz (durch die Verwendung der abgekürzten Form, d. h. eine Ablehnung der Regel ist denkbar) und der Eindeutigkeit (durch die Verwendung vollständiger Wörter, d. h. eine Anwendung der Regel ist zwingend erforderlich) individuell angewendet werden. Je kritischer der Kontext ist (z. B. in Sicherheitsanweisungen), desto mehr kann die Eindeutigkeit priorisiert werden.

Bei *„pas – Passiv vermeiden“* sanken nach der Regelanwendung systemübergreifend alle Qualitätsparameter (SQ, CQ und beide AEMs) signifikant. Die Regel „Passiv vermeiden“ ist eine weitverbreitete KS-Regel. Mehrere Studien argumentieren, dass das Vermeiden des Passivs die maschinelle Übersetzbarkeit verbessere, da so grammatische Parsing-Probleme umgangen werden könnten (vgl. Bernth & Gdaniec 2001; Reuther 2003; Fiederer & O'Brien 2009; Siegel 2013). Nach den Ergebnissen der Humanevaluation wurde die MÜ-Qualität des Passivs höher bewertet: Die Bewerter fanden das Aktiv (nach-KS) stilistisch nicht ideal für die Satzintention. In Bezug auf die Inhaltsqualität wurde die Genauigkeit des Passivs höher eingestuft. Auf Basis dieser Ergebnisse waren die Systeme in der Lage – im Gegensatz zu den vorherigen Studien, das Passiv inhaltlich und stilistisch mit einer hohen Qualität zu übersetzen. Das belegt einen Fortschritt bei der MÜ des Passivs und regt – anders als früher – dazu an, seine Verwendung in der Technikredaktion prinzipiell nicht zu verbannen, solange der Kontext und die Satzintention dies zulassen.

Parallel zu den triangulierten Ergebnissen der Fehlerannotation und der Humanevaluation (Tabelle 6.1) spiegelten die AEM-Scores (sowohl TERbase als auch hLEPOR) sowie die Humanscores (von SQ und CQ) wider, dass die Regeln „nsp – Eindeutige pronominale Bezüge verwenden“ und „prä – Überflüssige Präfixe vermeiden“ keinen signifikanten Einfluss auf die MÜ-Qualität hatten.

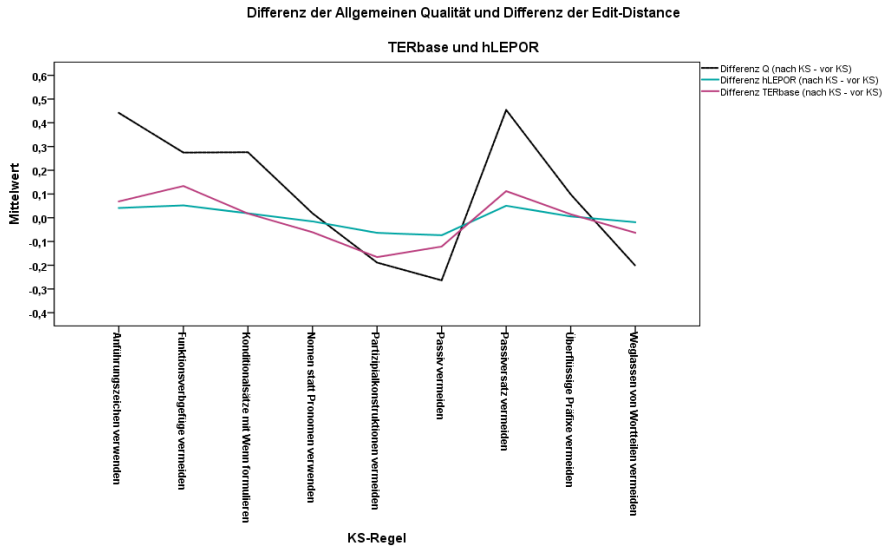
In Bezug auf *„nsp – Eindeutige pronominale Bezüge verwenden“*: Die Koreferenzauflösung[3] stellte bislang für die MÜ-Systeme eine Schwierigkeit dar (vgl. Ng 2017). Die Entscheidung, ein Pronomen zu verwenden oder es durch seine Referenz zu ersetzen, wird für gewöhnlich auf einer Fall-zu-Fall-Basis abhängig von dem Satz und der Formulierung der vorangehenden und folgenden Sätze getroffen (Bernth & Gdaniec 2001). So betrachteten Bernth und Gdaniec (ebd.: 187) das Vermeiden der Pronomen als einen „trade-off between MTranslatability and natural-sounding language“. Dies könnte der Grund sein, warum keine signifikante Auswirkung festgestellt werden konnte: Auf der einen Seite war die Verwendung der pronominalen Referenz von Vorteil, wenn die Identifizierung der Referenz als schwierig eingeschätzt wurde. Dies führte zu einer Erhöhung der MÜ-Klarheit, worauf tekom (2013: 137) und Congree (2018) ebenfalls abzielen. Auf der anderen Seite wurde die Wiederholung der Referenz in einigen Fällen auf stilistischer Ebene kritisiert. Wie Abbildung 6.2 zeigt, waren die MÜ-Systeme in 70 % der Fälle (Gruppe RR plus RF) in der Lage, die Koreferenzen korrekt aufzulösen. Dies stellt eine relativ hohe Erfolgsquote bei der Problematik der Koreferenzauflösung dar, wodurch ein natürlicher Stil gefördert wird. Dieser Fortschritt erlaubt es, die Regalanwendung auf die Fälle, bei denen keine Mehrdeutigkeit toleriert werden kann, z. B. kritischen Kontexte wie Sicherheitsanweisungen, einzuschränken.

In Bezug auf die Regel *„prä – Überflüssige Präfixe vermeiden“* war die RR-Annotationsgruppe sehr dominant (68 %). Wenn die MÜ-Systeme ein Verb mit und ohne überflüssiges Präfix (z. B. ‚anbieten‘ und ‚bieten‘) korrekt übersetzen konnten, waren die Übersetzungen in beiden Fällen identisch (richtige Übersetzung: ‚*offer*‘). Eine große Anzahl korrekter identischer Übersetzungen vor und nach der Regelanwendung führte entsprechend zu einer vergleichbaren Qualität. Das begründet das insignifikante Ergebnis der Qualitätsdifferenz (vor vs. nach der Regelanwendung). Gleichzeitig zeigt das Ergebnis der Gruppe FR, dass nach der Regelanwendung MÜ-Fehler in 13 % der Fälle – meistens bei Präfix-Verbgetrennten Fällen – behoben wurden und die Stil- und Inhaltsqualität dieser Gruppe signifikant anstiegen. Systemübergreifend stimmt in diesem Sinne das Ergebnis mit der tekom (2013: 111) sowie vorherigen Studien (Bernth & Gdaniec 2001; Siegel 2011; Siegel 2013) überein, dass diese Regel zur Förderung der maschinellen Übersetzbarkeit beiträgt.

Wie Abbildung 6.4 zeigt, hatten die aus der Humanevaluation resultierenden sowie die in der automatischen Evaluation errechneten Qualitätsveränderungen

[3] Durch die Koreferenzauflösung wird die Entität, auf die sich die Koreferenz bzw. das Pronomen bezieht, identifiziert (vgl. Ng 2017).

einen vergleichbaren Verlauf. Ein solcher vergleichbarer Verlauf deutet darauf hin, dass die Ergebnisse der beiden Methoden sich gegenseitig untermauern, denn die MÜ-Qualität wurde nach den beiden Methoden auf unterschiedliche Weise untersucht und gemessen.



Differenz = Szenario nach KS *minus* Szenario vor KS

Abbildung 6.4: Differenz der allgemeinen Qualität und Differenz der AEM-Scores auf Regelebene

Darüber hinaus zeigte der Spearman-Test für die Korrelation zwischen der Differenz in der allgemeinen Qualität und der Differenz in den AEM-Scores bei den einzelnen Regeln, dass die in der Humanevaluation und automatischen Evaluation festgestellten Qualitätsveränderungen übereinstimmen bzw. sich gegenseitig bestätigen: Bei den Regeln „fvg – Funktionsverbgefüge vermeiden“, „kos – Konditionalsätze mit ‚Wenn‘ einleiten“ und „pas – Passiv vermeiden“ gab es eine signifikante starke positive Korrelation ($\rho > 0{,}5$); bei den verbleibenden Regeln eine signifikante mittlere positive Korrelation ($\rho > 0{,}3$).

6.4 Auswirkung der KS auf Regel- und MÜ-Systemebene

Bisher zeigten die Ergebnisse auf Regelebene, dass vier Regeln einen positiven Einfluss auf die MÜ-Qualität haben („anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“, „per – Konstruktionen mit ‚sein + zu + Infinitiv‘

vermeiden“, „fvg – Funktionsverbgefüge vermeiden“ und „kos – Konditionalsätze mit ‚Wenn‘ einleiten“); und drei Regeln haben tendenziell einen negativen Einfluss auf die MÜ-Qualität („pak – Partizipialkonstruktion vermeiden“, „pas – Passiv vermeiden“ und „wte – Keine Wortteile weglassen“) – insbesondere hinsichtlich der Stilqualität. Für diese sieben Regeln wurde auf MÜ-Systemebene untersucht, welche Systeme den identifizierten Effekt aufweisen. Die Auswirkung der beiden verbleibenden Regeln „nsp – Eindeutige pronominale Bezüge verwenden“ und „prä – Überflüssige Präfixe vermeiden“ war nicht eindeutig. Für diese Regeln wurde auf MÜ-Systemebene genauer geprüft, ob bei einem bestimmten System signifikante Auswirkungen nachweisbar sind.

Bei der Diskussion werden die Systeme basierend auf den erzielten Ergebnissen verglichen. Die Qualitätsveränderungen werden in Zusammenhang mit den korrelierenden Fehlertypen (Ergebnisse der Fehlerannotation) bzw. den beeinflussten Qualitätskriterien (siehe [3a und 3b] in Abbildung 4.6) erörtert. Die Lieferung einer exakten Interpretation, was genau im Hintergrund jedes Systems zu einem bestimmten Output (bzw. zum Auftritt oder zur Aufhebung eines bestimmten Fehlers) geführt hat, erfordert eine Glas-Box-Analyse und geht somit über den Umfang dieser Studie hinaus.

Tabelle 6.2 liefert eine Übersicht über die sich positiv, negativ und nicht signifikant auswirkenden Regeln zusammen mit ihren Auswirkungen auf die Fehleranzahl und die Qualitätswerte bei den einzelnen Systemen:

6.4.1 Regeln mit positiver Wirkung

Regel *„anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“* war die einzige Regel, die mit einem Rückgang der Fehleranzahl sowie einer Verbesserung der SQ und CQ bei allen MÜ-Systemen verbunden war.

Das verbesserte Parsen, das durch die Kennzeichnung der Oberflächentexte mithilfe der Anführungszeichen realisiert wurde, war meist mit der Korrektur zweier Fehlertypen verbunden: Großschreibung des Oberflächentexts (OR.2) und Wortstellung (GR.10). Beim RBMÜ-System *Lucy* korrelierte die Korrektur dieser beiden Fehlertypen nach der Regelanwendung stark mit der Qualitätssteigerung. Im HMÜ-System *Bing* korrelierte nur die Korrektur des Wortstellungsfehlers stark mit der Qualitätsverbesserung. In den anderen Systemen konnten keine Korrelationen zwischen den Fehlertypen und der Qualität festgestellt werden.

Der Rückgang der Fehleranzahl nach der Regelanwendung war im Fall des anderen HMÜ-Systems *Systran* und des NMÜ-Systems *Google Translate* aus jeweils unterschiedlichen Gründen nicht signifikant. In *Systran* war die Fehleranzahl

Tabelle 6.2: Einfluss der einzelnen Regeln auf die Fehleranzahl sowie Stil- und Inhaltsqualität bei den einzelnen Systemen

		HMÜ (Bing)		NMÜ (Google)		RBMÜ (Lucy)		SMÜ (SDL)		HMÜ (Systran)	
-anz-	Anno.	F (- -)		F (-)		F (- -)		F (- -)		F (-)	
	HuEv	SQ ++	CQ ++	SQ ++	CQ +	SQ ++	CQ ++	SQ ++	CQ ++	SQ ++	CQ ++
-per-	Anno.	F (- -)		F =		F ++		F (- -)		F ++	
	HuEv	SQ ++	CQ ++	SQ +	CQ (-)	SQ +	CQ (-)	SQ ++	CQ ++	SQ ++	CQ (-)
-fvg-	Anno.	F (-)		F (-)		F (- -)		F (- -)		F (- -)	
	HuEv	SQ +	CQ +	SQ +	CQ (-)	SQ ++	CQ +	SQ +	CQ +	SQ ++	CQ +
-kos-	Anno.	F (- -)		F +		F (-)		F (- -)		F (-)	
	HuEv	SQ ++	CQ ++	SQ =	CQ +	SQ (-)	CQ (-)	SQ +	CQ ++	SQ (-)	CQ (-)
-pak-	Anno.	F +		F +		F +		F (-)		F ++	
	HuEv	SQ (- -)	CQ (-)	SQ (- -)	CQ (-)	SQ (- -)	CQ (-)	SQ (-)	CQ +	SQ (-)	CQ ++
-pas-	Anno.	F (- -)		F +		F +		F +		F ++	
	HuEv	SQ (-)	CQ +	SQ (-)	CQ (-)	SQ (- -)	CQ (-)	SQ (-)	CQ (-)	SQ (- -)	CQ (- -)
-wte-	Anno.	F +		F (-)		F (-)		F =		F (-)	
	HuEv	SQ (-)	CQ (-)	SQ (- -)	CQ (-)	SQ (- -)	CQ (-)	SQ (- -)	CQ (-)	SQ (-)	CQ +
-nsp-	Anno.	F =		F ++		F (-)		F (-)		F (-)	
	HuEv	SQ (-)	CQ +	SQ (- -)	CQ (-)	SQ +	CQ ++	SQ (-)	CQ +	SQ (-)	CQ (-)
-prä-	Anno.	F (- -)		F (-)		F (-)		F (-)		F (-)	
	HuEv	SQ +	CQ +	SQ (-)	CQ (-)	SQ +	CQ (-)	SQ +	CQ +	SQ +	CQ +

Regeln , die auf *Regelebene* einen signifikanten + Einfluss zeigten

Regeln , die auf *Regelebene* einen signifikanten (-) Einfluss zeigten

Regeln , die auf *Regelebene* keinen signifikanten Einfluss zeigten

Graue Zellen : signifikante Veränderung

F: Fehleranzahl gleich geblieben; SQ: Stilqualität; CQ: Inhaltsqualität; (- -) signifikanter Rückgang; ++ signifikanter Anstieg; (-) nicht signifikanter Rückgang; + nicht signifikanter Anstieg; = unverändert geblieben

sehr hoch und änderte sich kaum nach der Regelanwendung. In *Google Translate* hingegen fiel die Fehleranzahl sowohl vor- als auch nach-KS sehr gering aus. Das *NMÜ-System Google Translate* war in der Lage 83 % der Sätze mit und ohne Anführungszeichen fehlerfrei zu übersetzen (gefolgt von nur 17 % bei dem SMÜ-System *SDL*). Dementsprechend verzeichnete die NMÜ die höchsten SQ- und CQ-Werte.

Der Anstieg in der SQ bei dieser Regel war der einzige signifikante Anstieg beim NMÜ-System Google Translate unter allen Regeln (siehe Tabelle 6.2). Dieser Anstieg kam im Datensatz durch drei Veränderungen zustande: Die erste Veränderung beruhte darauf, dass der Oberflächentext vor KS als Terminus erkannt und großgeschrieben wurde, jedoch ohne Hervorhebung (z. B. *Change device parameters function* als Übersetzung für *Funktion Geräteparameter ändern*); mit der Anwendung der Regel wurde die MÜ zusätzlich mithilfe der Anführungszeichen hervorgehoben (*function "Change device parameters"*). Zweites gab es Fälle, bei denen der Oberflächentext als Terminus nicht erkannt und kleingeschrieben wurde (z. B. *additional information button* als Übersetzung für *Taste Zusatzinformation*); eine Erkennung des Terminus gelang nach der Regelanwendung und somit wurde er großgeschrieben und mithilfe der Anführungszeichen hervorgehoben ("*Additional information" button*). Die dritte Veränderung kam in zwei (von 24) Fällen vor, bei denen Oberflächentexte, die aus mehreren Wörtern bestanden (sog. Mehrwortentitäten), zerlegt wurden, was in Wortstellungsfehlern resultierte (z. B. *Upload function …from the device* als Übersetzung für *Funktion Upload vom Gerät*). Durch die Regelanwendung wurde der Wortstellungsfehler behoben ("*Upload from device" function*).

Auf Basis dieser Ergebnisse, die aus der NMÜ (Stand: Ende 2016) hervorgehen, konnte die Verwendung von geraden Anführungszeichen das generische NMÜ-System Google Translate dabei unterstützen, Mehrwortentitäten korrekt zu parsen, spezifische bzw. seltene Termini korrekt zu übersetzen und somit Wortstellungsfehler zu vermeiden. Daraufhin stieg – nach der Humanevaluation – die Stilqualität, i. S. v. Eignung der Übersetzung für die Intention ihres Inhalts (Hutchins & Somers 1992: 163) nach der Regelanwendung. Problematisch dabei blieb, dass die Hervorhebung keine typische Anwendung der Anführungszeichen im Englischen ist (McMurrey 2006). Dieser Einwand kam in den Kommentaren der Bewerter wiederholt vor. In dieser Phase der NMÜ-Entwicklung (2016 / 2017) thematisierten vorherige Studien die Schwäche der NMÜ bei der Übersetzung von seltenen Wörtern und Eigennamen (vgl. Le & Schuster 2016; Koehn 2017) sowie die Problematik der Terminologieintegration in der NMÜ (vgl. Eisold 2017; Koehn 2017).

Bei einer wiederholten Übersetzung des Datensatzes Anfang 2020 stieg der Prozentsatz der fehlerfreien Übersetzungen (Annotationsgruppe RR) von 83 % auf 96 % (konkret beinhaltete nur ein Satz von 24 Sätzen einen Großschreibungsfehler bei der Übersetzung *ohne* Anführungszeichen). Diese Verbesserung deutet auf einen Fortschritt bei der Übersetzung von seltenen Wörtern bzw. Eigennamen sowie bei der Identifizierung von Mehrwortentitäten (wie z. B. der Optionsbezeichnung ‚Auswahl nach Anwendungs-Code') hin. Ferner wurden zuletzt einige Ansätze zur Terminologieintegration in der NMÜ entwickelt, die einen Fortschritt bei der konsistenten Übersetzung spezifischer Termini (einschließlich der Out-of-vocabulary-Fälle)[4] nach festgelegten externen Terminologielisten (bzw. Terminologiedatenbanken) und eine damit verbundene Steigerung der Gesamtübersetzungsqualität belegen (vgl. Chatterjee u. a. 2017; Hasler u. a. 2018; Dinu u. a. 2019).[5] Weitere aktuelle Studien konnten mithilfe mehrerer Strategien einen Fortschritt bei der Übersetzung unterschiedlicher Mehrwortausdrücke (Multiword Expressions, MWEs) realisieren (vgl. Gamallo & Garcia 2019; Rikters & Bojar 2019; Zaninello & Birch 2020).

Auf Basis dieser Entwicklung würde das Terminologiemanagement das System bei der Identifizierung und Übersetzung der spezifischen Termini und Mehrwortentitäten unterstützen. Gleichzeitig kann eine vom Unternehmen bestimmte Formatierung (z. B. kursiv) für die Hervorhebung sorgen, um die Eignung der Übersetzung im Sinne der Textintention sicherzustellen. Mit dieser Konstellation wäre die Verwendung der geraden Anführungszeichen bei Oberflächentexten zur Unterstützung der maschinellen Übersetzbarkeit nicht mehr erforderlich.

Für die zweite Regel *„fvg – Funktionsverbgefüge vermeiden"* war der allgemeine positive Einfluss auf den MÜ-Output auf Systemebene wie folgt:

Für das RBMÜ-System *Lucy* und ein HMÜ-System (*Systran*) war diese Regel besonders vorteilhaft, denn nach ihrer Anwendung sank die Fehleranzahl und die SQ verbesserte sich signifikant. In dem anderen HMÜ-System (*Bing*) und dem SMÜ-System *SDL* nahm die Fehleranzahl ab und die SQ und CQ nahmen zu; die Änderungen waren jedoch nicht signifikant. Die Bewerter stellten fest, dass die Verwendung des bedeutungstragenden Verbs (nach-KS) anstelle des Funktionsverbgefüges (vor-KS) die Übersetzung verständlicher und stilistisch aufmerksamkeitserregend ausfallen lässt, was in Einklang mit der Empfehlung der Regelsätze von Congree (2018) und Siegel (2011) zur Regelanwendung steht. Dieses Ergebnis deutet ferner darauf hin, dass der Effekt, der von der tekom (2013: 107) bei der Ausgangssprache, beabsichtigt ist, nämlich den Satz konkreter und direkter zu gestalten, sich in der Zielsprache widerspiegelt. Die Analyse des *NMÜ-Systems Google Translate* ergab distinkte Ergebnisse: Die Fehleranzahl war minimal; Google

[4] „Out of vocabulary" sind externe Wörter, die dem Modell unbekannt sind (vgl. Eisold 2017).
[5] Mehr zu diesen Studien unter §3.3.4.

Translate war in der Lage, 88 % der Sätze vor und nach der Regelanwendung fehlerfrei (Gruppe RR) zu übersetzen (gefolgt von 46 % in Bing). Es verzeichnete die höchsten SQ und CQ unter allen Systemen sowohl vor als auch nach der Regelimplementierung.

Da nicht alle deutschen Funktionsverbgefüge ein Pendant auf Englisch haben, waren die ausdrucksschwachen Verben für die Systeme oft ambig bzw. es entstand ein Transferproblem (vgl. Baumert & Verhein-Jarren 2012: 107). Das „Vermeiden des Funktionsverbgefüges" hat überwiegend die Satzsemantik positiv beeinflusst. Die Verwendung des bedeutungstragenden Verbs anstelle des Funktionsverbgefüges war mit der Korrektur einer Reihe semantischer Fehler, insbesondere Kollokationsfehler (SM.13), sowie lexikalischer Fehler verbunden. Lexikalische Fehler traten dann auf, wenn die Systeme das Funktionsverbgefüge wörtlich übersetzten (z. B. das Übersetzen von ‚zur Verfügung stellen' als ‚represent available' statt ‚provide'). In *Lucy* korrelierte die Korrektur der semantischen Fehler mit einem Anstieg von SQ und CQ. In *Bing* und *SDL* war eine Korrelation zwischen den lexikalischen Fehlern *Wort ausgelassen* (LX.3) und Wort zusätzlich falsch eingefügt (LX.4) und der Qualität zu beobachten. In den anderen Systemen waren keine weiteren Korrelationen nachweisbar.

Wie unter §4.5.2.3 diskutiert, ist die Verwendung des Funktionsverbgefüges in manchen Fällen zum Ausdrücken von bestimmten Bedeutungsnuancen oder mangels eines äquivalenten bedeutungstragenden Verbs erforderlich (Baumert & Verhein-Jarren 2012: 107). Auf Basis der Ergebnisse bietet der NMÜ-Ansatz im Vergleich zu den früheren Ansätzen in solchen Fällen eine solide Architektur zur Übersetzung des Funktionsverbgefüges. Somit ist das Vermeiden des Funktionsverbgefüges, wie es von Regelsätzen wie Congree (2018) und Siegel (2011) empfohlen ist, zum Zwecke der maschinellen Übersetzbarkeit nicht erforderlich.

Die Anwendung der Regel *„kos – Konditionalsätze mit ‚Wenn' einleiten"* war mit einer Verringerung der Fehleranzahl in allen Systemen mit Ausnahme des *NMÜ-Systems Google Translate* (aufgrund seiner geringen Fehleranzahl) verbunden. Das Einleiten von Konditionalsätzen mit ‚Wenn' zeigte hauptsächlich eine positive lexikalische Wirkung, die oft mit der Korrektur eines Wortstellungsfehlers verbunden war. Während deutsche Konditionalsätze mit einem Verb am Satzbeginn (d. h. ohne die Konditionalkonjunktion ‚Wenn') formuliert werden können, ist dies im Englischen nicht der Fall. Daher war das Weglassen der Konjunktion (vor KS) mit zwei Fehlertypen verbunden: dem Fehlen der Konjunktion ‚Wenn' (LX.3) und der falschen Platzierung des Verbs des Konditionalsatzes (GR.10). Diese Fehler sanken nach der Regelanwendung und ihr Rückgang korrelierte entsprechend negativ mit der Qualitätssteigerung, und zwar eine starke Korrelation im Falle des LX.3 bzw. eine mittlere Korrelation im Falle des GR.10.

Der MÜ-Output der einzelnen Systeme wurde unterschiedlich beeinflusst. Das zeigte wiederum, dass die Regelanwendung nicht bei allen Systemen erforderlich war. Im Folgenden eine genaue Untersuchung auf Systemebene: Der Rückgang der MÜ-Fehler war nur bei einem HMÜ-System (*Bing*) und dem SMÜ-System *SDL* signifikant: sowohl LX.3 als auch GR.10 bei Bing und nur LX.3 bei SDL. Infolgedessen wurde nur in diesen beiden Systemen eine signifikante Qualitätsverbesserung erzielt – in Bing sowohl für die SQ als auch für die CQ und in SDL nur für die SQ. In *Bing* und *SDL* korrelierte ebenfalls die Korrektur dieser Fehler stark mit der Qualitätssteigerung. In den anderen Systemen wurden keine Korrelationen mit einem bestimmten Fehlertyp beobachtet. Nach der Regelanwendung fanden die Bewerter die MÜ genauer, verständlicher und idiomatischer. Im Gegensatz dazu sanken beide Qualitätswerte im RBMÜ-System *Lucy* und im anderen HMÜ-System (*Systran*) nach der Regelanwendung. Das liegt daran, dass Lucy 71 % und Systran 58 % vor und nach der Regelanwendung fehlerfrei übersetzen konnten (Gruppe RR). Bei den beiden Systemen wurden die Fehler nur in zwei Fällen nach der Regelanwendung behoben (Gruppe FR). Dementsprechend zeigen Lucy und Systran einen gewissen Fortschritt bei der Übersetzung dieser Art der elliptischen Konstruktion. In *Google Translate* betrug der Prozentsatz fehlerfreier MÜ vor und nach der Regelanwendung (Gruppe RR) 92 % (gefolgt von 71 % in Lucy). Dies zeigte erneut die höchsten SQ und CQ in beiden Szenarien und weist darauf hin, dass das NMÜ-System die sprachlichen Unterschiede sowie Syntaxanalyse dieser elliptischen Konstruktion meistern konnte.

Regel *„per – Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden"* zeigte eine generell positive Auswirkung auf den MÜ-Output:

In Einklang mit Reuther's (2003) Beobachtung, dass die systemspezifischen Eigenschaften Einfluss auf die maschinelle Übersetzung stilistischer Phänomene wie z. B. der Passiversatzkonstruktion (sein + zu + Infinitiv) haben, wurden (vor-KS) grammatische Schwierigkeiten nur bei zwei Systemen, dem HMÜ-System *Bing* und dem SMÜ-System *SDL*, beobachtet. Nach der Regelanwendung sanken zwei Grammatikfehler signifikant: GR.8 falsches Verb (Zeitform, Komposition, Person) und GR.10 falsche Wortstellung. In *Bing* und *SDL* korrelierte die Korrektur dieser Fehlertypen stark mit einem Anstieg der Qualitätswerte. Auf der anderen Seite war die Regelanwendung mithilfe des Imperativs (anstelle der Konstruktion „sein + zu + Infinitiv") mit dem Auftreten des lexikalischen Additionsfehlers (LX.4) (nach KS) im RBMÜ-System *Lucy* und im anderen HMÜ-System (*Systran*) verbunden, da die MÜ-Systeme in einigen Fällen fälschlicherweise das Subjekt *you* hinzufügten (z. B. ‚*Verriegeln Sie* die Kontakte' → ‚*You lock* the contacts'). In *Systran* korrelierte die Korrektur dieses lexikalischen Fehlers stark und signifikant mit dem Anstieg der Qualitätswerte. Das *NMÜ-System Google Translate* konnte 96 % der Sätze vor und nach der Regelanwendung fehlerfrei (Gruppe

RR) übersetzen (gefolgt von nur 42 % in Systran). Daher war die Stil- und Inhaltsqualität des MÜ-Outputs in beiden Szenarien bei dem NMÜ-System Google Translate am höchsten unter allen MÜ-Systemen (mit einer minimalen Zunahme der SQ und einer minimalen Abnahme der CQ nach KS).

6.4.2 Regeln mit negativer Wirkung

Die Regel *„pak – Partizipialkonstruktion vermeiden"* wirkte sich in allen MÜ-Systemen generell negativ auf die SQ aus. Die CQ (überwiegend die Verständlichkeit) stieg nur in zwei MÜ-Systemen an: minimal im SMÜ-System *SDL* und signifikant in einem HMÜ-System (*Systran*). Die negative Wirkung war auch in Bezug auf die Fehleranzahl zu beobachten. Die Fehleranzahl stieg nach der Regelanwendung in allen Systemen an, außer in *SDL*, wo ein leichter Rückgang festgestellt wurde. Bei *Systran* fiel der Anstieg der Fehleranzahl signifikant aus.

Das *NMÜ-System Google Translate* hatte keine Schwierigkeit, Partizipialkonstruktionen zu übersetzen. Ferner waren 71 % der Übersetzungen von *Google Translate* vor und nach der Regelanwendung fehlerfrei, d. h. Annotationsgruppe RR, (gefolgt von nur 29 % in Bing). In allen anderen Systemen war die Fehleranzahl sowohl vor als auch nach der Regelimplementierung wesentlich höher. Darüber hinaus zeigte Google Translate die höchsten Qualitätswerte in beiden Szenarien, was seinen Fortschritt beim Parsen von komplexen Strukturen wie Partizipialkonstruktionen widerspiegelt.

Mit dieser Regel waren zwei verschiedene MÜ-Fehlertypen verbunden: Erstens erschweren besonders lange Partizipialkonstruktionen die Satzstruktur und damit das Parsen, was zu Wortstellungsfehlern (GR.10) führte. Dieser Fehlertyp trat insbesondere beim SMÜ-System *SDL* und HMÜ-System *Bing* bei der Übersetzung von Partizipialkonstruktionen auf und wurde nach der Regelanwendung behoben. Nach der Implementierung der Regel hatten jedoch alle Systeme Probleme mit der Platzierung von Kommas in Nebensätzen, insbesondere in denen eine Unterscheidung zwischen ‚which' und ‚that' gemacht werden musste (vgl. Swan 1980: 527 ff.).[6] Daher stieg die Anzahl der Zeichensetzungsfehler (OR.1) bei Bing, Systran und SDL signifikant an. Im Gegensatz zum Deutschen werden im Englischen nach der Regelanwendung für den Nebensatz nicht immer Kommas benötigt. Nichtsdestotrotz ist es zu beachten, dass die Verwendung von ‚which'

[6] Man unterscheidet im Englischen zwischen restriktiven und nicht-restriktiven Relativsätzen: Im restriktiven Relativsatz wird die Bedeutung des Nomens, das beschrieben wird, begrenzt. Ohne den restriktiven Relativsatz ändert sich die Bedeutung des gesamten Satzes. Für restriktive Relativsätze verwendet man ‚that'. Der nicht-restriktiver Relativsatz bietet lediglich zusätzliche Informationen über das Nomen und kann ohne Einfluss auf die Bedeutung entfernt werden. Für nicht-restriktive Relativsätze verwendet man ‚which'. (vgl. Swan 1980: 527 ff.)

vs. ‚that' im Allgemeinen nicht selten problematisch ist und in der Regel kontextbezogene Informationen erfordert, um über die korrekte Verwendung entscheiden zu können (vgl. ebd.).

Bei der Regel *„pas – Passiv vermeiden"* zeigten alle MÜ-Systeme – mit Ausnahme von Bing – nach der Regelanwendung einen Anstieg der Fehleranzahl. Die beiden HMÜ-Systeme lieferten gegensätzliche Ergebnisse: Bei *Bing* verringerte sich die Gesamtzahl der Fehler signifikant, während sie bei *Systran* erheblich zunahm. In den anderen drei Systemen war die Zunahme der Fehleranzahl nicht signifikant. *Bing* und *Google Translate* konnten 71 % der Sätze sowohl im Passiv als auch im Aktiv fehlerfrei (Gruppe RR) übersetzen (gefolgt von 58 % in Lucy).

Generell unterstreicht das Ergebnis den Fortschritt bei der MÜ des Passivs. Nach der Regelanwendung sank sowohl die SQ als auch die CQ in allen MÜ-Systemen, mit Ausnahme von *Bing*, in dem die CQ leicht anstieg. In *Systran* war der SQ- und CQ-Rückgang signifikant. Bei *Lucy* war der Rückgang der SQ signifikant. Sowohl vor als auch nach der Implementierung der Regel lieferte *Google Translate* die höchsten SQ und CQ. Die Qualitätsveränderungen korrelierten bei keinem System mit einem bestimmten Fehlertyp, denn die Regelanwendung war mit einer Zunahme verschiedener Fehlertypen bei allen Systemen verbunden. Diese Zunahme (nach KS) war allerdings bei keinem der Fehlertypen signifikant.

Regel *„wte – Keine Wortteile weglassen"*: Der Ergänzungsstrich signalisiert die Auslassung eines Wortteils sowie die „Zusammengehörigkeit räumlich getrennter Bestandteile zusammengesetzter oder abgeleiteter koordinierter Wörter innerhalb der Wortgruppe" (Nerius 2007: 191). Gleichzeitig spielen die Satzzeichen wie in diesem Fall der Ergänzungsstrich laut Reuther (2003: 2) eine wesentliche Rolle bei der MÜ: „Punctuation marks are very sensitive with respect to all applications where linguistic processing is done automatically." Diese Regel war mit einer geringen Abnahme der Fehleranzahl bei *Google Translate*, *Lucy* und *Systran* verbunden. Aufgrund der Unterschiede in den Rechtschreibregeln im Deutschen vs. Englischen in Bezug auf die Verwendung des Ergänzungsstriches trug die Regel zu einer verbesserten Tokenisierung bei. So wurde beobachtet, dass die Regelanwendung mit einer Verringerung des Zeichensetzungsfehlers verbunden war. Die Fehleranzahl stieg jedoch im HMÜ-System *Bing* marginal an und blieb im SMÜ-System *SDL* unverändert.

Trotz der unterschiedlichen Auswirkungen auf die Fehleranzahl sanken SQ und CQ in allen Systemen, mit Ausnahme von *Systran*, bei dem die CQ leicht anstieg. Leichte Anstiege in der CQ kamen bei Wortkonstellationen vor, die in ihrer abgekürzten Form nicht gebräuchlich sind (z. B. ‚Wasser-, Gasrohre oder stromführende Leitungen' (im Vergleich zu für gewöhnlich abgekürzten Begriffen wie

‚Vor- und Nachteile‘)). Bei ungeläufigen abgekürzten Begriffen und insbesondere bei einem kritischen Inhalt, z. B. bei wichtigen Warn- oder Fehlermeldungen einer Maschine, verbesserte die Regel die Eindeutigkeit. Die SQ-Abnahme war in drei Systemen (*Google Translate*, *Lucy* und *SDL*) signifikant, da die Bewerter die Nomenwiederholung als unnatürlich empfanden (z. B. *Milch* in *Sojamilch und laktosefreie Milch* nach-KS anstelle von *Soja- und laktosefreie Milch* vor-KS) (siehe „Quantitätsmaxime“ von Grice 1975: 26).[7] Auch bei dieser Regel zeigte *Google Translate* die geringste Fehleranzahl (75 % der Übersetzungen waren sowohl vor als auch nach KS fehlerfrei, gefolgt von 46 % in SDL (Gruppe RR)) und die höchsten Qualitätswerte in beiden Szenarien.

Was die Übersetzbarkeit betrifft, zeigte die Regel „*Keine Wortteile weglassen*“ in Bezug auf die Fehleranzahl, die Qualitätsratings und die AEM-Scores keine signifikante Verbesserung. Obwohl die Ergebnisse der Studie von einer vorangestellten Terminologieintegration sicherlich positiv beeinflusst worden wären, konnte Google Translate auch ohne Terminologieintegration die Wortteile in den meisten Fällen korrekt übersetzen.

6.4.3 Regeln ohne signifikante Auswirkung

Die Regel *„nsp – Eindeutige pronominale Bezüge verwenden“* hatte von einem System zum anderen unterschiedliche Auswirkungen auf die MÜ-Qualität. Nur das RBMÜ-System *Lucy* und das *NMÜ-System Google Translate* zeigten signifikante Qualitätsveränderungen: *Lucy* zeigte eine leichte Zunahme der SQ und eine signifikante Zunahme der CQ, während für *Google Translate* genau das Gegenteil der Fall war – nämlich eine signifikante Abnahme der SQ und ein leichter Rückgang der CQ. Dies könnte durch die verschiedenen Änderungen in der Fehleranzahl erklärt werden. Im Gegensatz zu den anderen Systemen konnte *Google Translate* die Pronomen meist korrekt übersetzen (vor-KS), während die Verwendung einer pronominalen Referenz (nach-KS) in einigen Fällen stilistisch kritisiert wurde. Gleichzeitig waren 83 % der Übersetzungen von *Google Translate* vor und nach der Regelanwendung fehlerfrei (gefolgt von 67 % in *Lucy*). Darüber hinaus erzielte *Google Translate* in beiden Szenarien die höchsten Qualität-Scores.

Die Verwendung eindeutiger pronominaler Referenzen zeigte zwei unterschiedliche Auswirkungen auf die MÜ-Systeme: In *Lucy*, *SDL* und *Systran* war die Regelanwendung mit einer Reduzierung des semantischen Fehlers Verwechselung

[7] Die Quantitätsmaxime von Grice (1975: 26) bezieht sich auf die Quantität der Informationen: „Make your contribution as informative as is required (for the current purpose of the exchange) […] not more informative than is required”.

des Sinns (SM.11) verbunden. Dieser Fehler trat insbesondere bei der Übersetzung von Demonstrativpronomen (‚diese' und ‚dies') auf, da die MÜ-Systeme Schwierigkeiten hatten, die Referenz zu identifizieren und korrekt zu übersetzen. Dementsprechend fanden die Bewerter die Übersetzung nach der Anwendung der Regel eindeutiger. Auf der anderen Seite war die Anwendung dieser Regel mit einem Anstieg des lexikalischen Konsistenzfehlers (LX.6) in *Bing* und *SDL* verbunden. Um diese Regel zu implementieren, sollte ein Substantiv im Hauptsatz *nicht* durch ein Pronomen im Nebensatz ersetzt werden; stattdessen sollte der Pronomenbezug verwendet werden. In einigen Fällen übersetzten die MÜ-Systeme die zweite Instanz des Substantivs anders (bzw. verwendeten Synonyme), was zu einem Konsistenzfehler (vgl. Mertin 2006: 249) und damit zu einer verringerten Genauigkeit führte. Ein Konsistenzfehler könnte jedoch vermieden werden, wenn die verwendeten Termini in die Systeme integriert würden.[8]

Regel *„prä – Überflüssige Präfixe vermeiden"* war – im Allgemeinen – mit einer sehr geringen Fehleranzahl vor der Regelanwendung verbunden, die nach der Regelanwendung abnahm. Diese Reduktion war nur in einem HMÜ-System (*Bing*) signifikant. Die SQ hat sich in allen Systemen mit Ausnahme von *Google Translate* leicht verbessert. Auch die CQ stieg in *Bing*, *Lucy* und *Systran* minimal an. Bei *Google Translate* wurden 88 % der Sätze vor und nach der Regelanwendung fehlerfrei übersetzt, d. h. Annotatiotionsgruppe RR, (gefolgt von 71 % bei Bing). Die Qualitätswerte in *Google Translate* zeigten nach der Regelanwendung eine minimale Abnahme. Gleichzeitig waren sie sowohl vor als auch nach der Implementierung der Regel die höchsten unter allen MÜ-Systemen.

Systemspezifisch kann das Ergebnis daher nur bis zu einem gewissen Grad die Empfehlung der tekom (2013: 111) sowie vorheriger Studien (Bernth & Gdaniec 2001; Siegel 2011; Siegel 2013) zur Regelanwendung im Hinblick auf die maschinelle Übersetzbarkeit bestätigen. Das Vermeiden überflüssiger Präfixe unterstützte alle Systeme mit Ausnahme des NMÜ-Systems Google Translate dabei, das Verb korrekt zu parsen. Insbesondere trennbare Verben waren oft schwer zu parsen; abhängig von der Satzstruktur stehen die Präfixe in manchen Fällen am Ende des Satzes – weit entfernt vom Rest des Verbs. In solchen Fällen übersetzten die Systeme das Präfix zusätzlich, d. h. unabhängig vom Verb. Die Anwendung der Regel führte zur Korrektur des lexikalischen Additionsfehlers (LX.4), was wiederum die Genauigkeit der Übersetzung verbesserte.

Wie die Ergebnisse der einzelnen Regeln auf Systemebene zeigen, korrelierten die Veränderungen der Fehlertypen mit denen der Qualitätswerte bei den älteren

[8] Die Studie wurde mit generischen Black-Box-Systemen durchgeführt. Für die Auswahlkriterien der Systeme siehe §4.5.1. Für den Umgang mit den spezifischen Termini im Rahmen der Studie siehe Schritt [4] unter §4.5.3.1.

Systemen (RBMÜ, SMÜ und HMÜ), während bei dem NMÜ-System gar keine einschlägige Korrelation bestand (Tabelle 6.4).

6.5 Regelübergreifende Auswirkung der KS auf MÜ-Systemebene

Bei der Diskussion der Ergebnisse auf Systemebene (regelübergreifend) werden die Systeme im Hinblick auf die MÜ-Qualitätsveränderungen verglichen, dabei werden die Qualitätsveränderungen in Zusammenhang mit den korrelierenden Fehlertypen (Ergebnisse der Fehlerannotation) sowie den beeinflussten Qualitätskriterien (siehe [3a und 3b] Abbildung 4.6) erörtert.

Vor der Diskussion der Ergebnisse auf Systemebene sind zwei Aspekte anzumerken: Erstens, die Studie wurde mithilfe generischer Black-Box-Systeme durchgeführt (für die Auswahlkriterien der Systeme siehe §4.5.1). Dementsprechend wurden die untersuchten Systeme nicht mit Korpora trainiert, die auf das Testmaterial abgestimmt sind. Ein Training der Systeme hätte einen Einfluss auf die Ergebnisse. Vorherige Studien (vgl. Ramírez Polo & Haller 2005; Aikawa u. a. 2007; Lehmann u. a. 2012) untersuchten diesen Einfluss. Bei der Arbeit mit trainierten Systemen werden erwartungsgemäß Phrasen bzw. Redewendungen, die in den Trainingsdaten vorkommen, korrekt übersetzt (Ramírez Polo & Haller 2005; Aikawa u. a. 2007). Gleichermaßen sollte eine bestimmte KS-Regel in den Trainingsdaten häufig angewendet werden, so übersetzt das System Segmente, die nach dieser Regel formuliert sind, besser als nicht-Regel-konforme Segmente (vgl. Lehmann u. a. 2012). Auf dieser Grundlage beabsichtigten Lehmann u. a. (2012) die Untersuchung einer automatischen Auswahl der anzuwendenden Regeln basierend auf dem verwendeten Trainingskorpus. In der vorliegenden Studie wurde hingegen gezielt mit Systemen gearbeitet, die vor der Untersuchung nicht mit bestimmten Korpora trainiert wurden. Würden die Systeme in der Studie vorab trainiert, wären die Ergebnisse von den Trainingsdaten abhängig (d. h. bessere Ergebnisse beim Kontrollierten Szenario, wenn die Korpora kontrolliert sind, und umgekehrt) (vgl. Reuther 2003). Außerdem wäre ein angemessener Vergleich der Ergebnisse der verschiedenen Systeme nicht realisierbar, denn die Systeme, die basierend auf Trainingskorpora arbeiten, hätten je nach Kontrollgrad einen Vorteil bzw. Nachteil gegenüber dem regelbasierten System.

Ferner wurde die Problematik der Terminologieübersetzung in der Studie umgangen, indem die spezifischen Termini in den analysierten Sätzen durch geläufige Begriffe ersetzt wurden (Genaueres dazu unter §4.5.3.1, Schritt [4]). Die Terminologieintegration wäre in der RMBÜ, SMÜ und HMÜ mithilfe fachspezifischer

Wörterbücher bzw. durch das Training mittels fachspezifischer Parallelkorpora möglich gewesen, jedoch befand sich die Terminologieintegration in der NMÜ zur Zeit der Durchführung der Studie noch in der experimentellen Phase (vgl. Eisold 2017). Damit die Studie auf einer einheitlichen Basis durchgeführt wird, wurden alle Systeme in ihrem Ist-Zustand, d. h. ohne Terminologieintegration oder Training mit domänenspezifischen Daten, verwendet.

Zweitens, jedes der analysierten Systeme hat je nach seinem Ansatz seine(n) eigene(n) Aufbau, Modelle, Abläufe, Trainingsdaten bzw. Lexika. Alle diese Komponenten beeinflussen den Output. Diese Komponenten bleiben aber bei der Analyse der Vor- und Nach-KS-Szenarien konstant. Die einzige Variable ist die Anwendung bzw. Nicht-Anwendung einer Regel. Da die Studienfrage den Fokus darauf legt, was die Anwendung der einzelnen Regeln bei den unterschiedlichen analysierten MÜ-Ansätzen bewirkt (und nicht warum bzw. wie eine Regel eine bestimmte Wirkung bei einem bestimmten MÜ-Ansatz zeigt), konnte die Frage im Rahmen einer Black-Box-Analyse beantwortet werden. Eine Untersuchung des Hintergrunds der Auswirkung (z. B. Grund des Auftritts oder der Aufhebung eines bestimmten Fehlers) bei jedem Ansatz erfordert eine Glas-Box-Analyse und geht somit über den Umfang dieser Studie hinaus.

Auf Systemebene zeigen die Ergebnisse, dass sich das getestete NMÜ-System bei Anwendung der KS-Regeln anders verhielt. Im Allgemeinen kann beobachtet werden, dass die analysierten KS-Regeln die Leistung der RBMÜ-, SMÜ- und HMÜ-Systeme verbesserten (Tabelle 6.3). Die Qualität wurde höher bewertet und die Fehleranzahl sank bei der Übersetzung der KS-Regel-konformen Ausgangssätze. Dies ist der Fall mit Ausnahme von Systran, dessen Fehleranzahl sowohl vor-KS als auch nach-KS sehr hoch war und nach-KS insignifikant stieg, mehr dazu siehe §5.5.2. Der NMÜ-Output wurde hingegen durch die Regeln nicht positiv beeinflusst; vor der KS-Anwendung fiel die Fehleranzahl geringer aus und die Qualität wurde höher bewertet.

Bezüglich der Fehleranzahl und der Fehlerannotationsgruppen zeigte das NMÜ-System die beste Leistung, unabhängig davon, ob die KS-Regeln angewendet wurden oder nicht (d. h. vor und nach KS), siehe §5.5.2 und §5.5.4. Bei einigen Regeln erhöhte sich sogar die Fehleranzahl in der KS-Stelle nach der Regelanwendung, siehe §5.4.5.2 (Regel 4) und §5.4.6.2 (Regel 5).

In Bezug auf die Fehlertypen zeigt Tabelle 6.4 die Fehlertypen, die bei jedem MÜ-System durch die Anwendung der KS-Regeln negativ oder positiv signifikant beeinflusst wurden:

Auf den ersten Blick erkennt man, dass es beim NMÜ-System im Gegensatz zu den anderen vier Systemen keine Fehlertypen gab, die von der KS-Anwendung

Tabelle 6.3: Reihenfolge der Systeme nach dem positiven Einfluss der KS-Regeln

	Fehleranzahl		**Stilqualität**		**Inhaltsqualität**
(1)	Bing (HMÜ) − 52,4 % signifikant	(1)	Bing + 5,2 % signifikant	(1)	Bing + 10,3 % signifikant
(2)	SDL (SMÜ) − 38,0 % signifikant	(2)	SDL + 5,0 % signifikant	(2)	SDL + 6,3 % signifikant
(3)	Lucy (RBMÜ) − 16,1 % *nicht* signifikant	(3)	Lucy + 1,0 % *nicht* signifikant	(3)	Systran + 1,1 % *nicht* signifikant
(4)	Google (NMÜ) + 6,1 % *nicht* signifikant	(4)	Systran + 0,7 % *nicht* signifikant	(4)	Lucy + 0,5 % *nicht* signifikant
(5)	Systran (HMÜ) + 10,1 % *nicht* signifikant	(5)	Google − 1,7 % signifikant	(5)	Google − 1,1 % *nicht* signifikant

signifikant beeinflusst wurden. Gefolgt vom NMÜ-System Google Translate zeigte das RBMÜ-System Lucy signifikante Änderungen in der Fehleranzahl von nur zwei Fehlertypen (OR.2 „Großschreibung“ und SM.13 „Kollokationsfehler“). Die höchste Anzahl von Fehlertypen, die von der Anwendung der Regeln signifikant betroffen waren, fand sich beim SMÜ-System SDL (6 Fehlertypen). Im Allgemeinen stehen diese Ergebnisse in Einklang mit denen einer vorherigen Studie (vgl. Lommel u. a. 2014), die zeigten, dass bei RBMÜ-Systemen semantische Fehler am häufigsten und Auslassungsfehler selten auftreten; und umgekehrt treten bei SMÜ-Systemen Auslassungs- und Wortstellungsfehler am häufigsten und semantische Fehler selten auf. Die Ergebnisse zeigen damit, dass die KS-Regeln bei der Beseitigung häufiger Fehlertypen beider Systemansätze von Nutzen waren. Schließlich verdeutlicht ein Vergleich der signifikanten Fehlertypen in den HMÜ-Systemen Bing und Systran, dass die Fehlertypen sehr unterschiedlich waren, was verschiedene Schwächen in den Architekturen beider HMÜ-Systeme widerspiegelt.

Bei dem NMÜ-System zeigt eine genauere Betrachtung der Regeln und der Fehlertypen, bei denen die Fehleranzahl zunahm, dass diese Zunahme hauptsächlich

Tabelle 6.4: Fehlertypen mit signifikanter Veränderung nach der KS-Anwendung auf MÜ-Systemebene

	Bing (HMÜ)	**SDL (SMÜ)**	**Lucy (RBMÜ)**	**Systran (HMÜ)**	**Google (NMÜ)**
OR.1 – Zeichensetzung	+	+		+	
OR.2 – Großschreibung	(–)	(–)	(–)		
LX.3 – Wort ausgelassen	(–)	(–)		(–)	
LX.4 – Wort zusätzl. falsch eingefügt		(–)		+	
LX.5 – Wort unübers. geblieben					
LX.6 – Konsistenzfehler					
GR.7 – Falsche Wortart					
GR.8 – Falsches Verb	(–)				
GR.9 – Kongruenzfehler		(–)			
GR.10 – Falsche Wortstellung	(–)	(–)			
SM.11 – Verwechsel. des Sinns					
SM.12 – Falsche Wahl					
SM.13 –Kollokationsfehler			(–)	(–)	

in Regel 4 „Eindeutige pronominale Bezüge verwenden“ (+4 Fehler) und Regel 5 „Partizipialkonstruktionen vermeiden“ (+6 Fehler in lediglich 4 Sätze) vorkam. In Bezug auf Regel 4 konnte das NMÜ-System die Pronomen größtenteils fehlerfrei übersetzen. Die Verwendung des Bezugsworts anstelle des Pronomens (d. h. nach der Regelanwendung) war in einigen Fällen mit einem Kongruenzfehler (Fehlertyp GR.9) verbunden. In Bezug auf Regel 5 hatte das NMÜ-System keine Schwierigkeiten, Partizipialkonstruktionen zu übersetzen und somit zeigte es eine hohe Leistung vor der Regelanwendung. Wenn das System jedoch einen Nebensatz anstelle der Partizipialkonstruktion (d. h. nach der Regelanwendung) übersetzte, trat der Fehlertyp OR.1 „Orthografie – Zeichensetzung“ auf, insbesondere in Fällen, in denen eine Unterscheidung zwischen den Relativpronomen ‚which‘ und ‚that‘ gemacht werden musste (vgl. Swan 1980: 527 ff.).[9] Dessen ungeachtet, wie die Kommentare bzw. vorgeschlagenen Übersetzungen der Humanevaluatoren zeigten, ist die Entscheidung über ‚which‘ vs. ‚that‘ in manchen Fällen nicht unkompliziert und normalerweise sind kontextbezogene Informationen für eine korrekte Verwendungsentscheidung erforderlich. Ein weiterer Fehlertyp, der im NMÜ-Output nach Anwendung der KS-Regeln anstieg, war der Fehlertyp SM.11 „Semantik – Verwechslung des Sinns“. Dieser Fehler trat jedoch in deutlich ambigen Fällen auf, wie z. B. bei der Übersetzung von ‚Wenn‘ als ‚When‘ anstelle von ‚If‘ in Regel 3 „Konditionalsätze mit ‚Wenn‘ einleiten“. Eine solche Sinnverwechslung wurde allerdings auch in den Ergebnissen der Humanevaluation beobachtet, da die Bewerter Kontextinformationen brauchten, um den Sinn in solchen Fällen zu klären, wie von den Bewertern kommentiert wurde.

Die Abnahme der Qualität nach der KS-Anwendung bei dem NMÜ-System war hauptsächlich auf die signifikant verringerte Stilqualität zurückzuführen. Die Inhaltsqualität ging zwar nach der KS-Anwendung ebenfalls zurück, jedoch war der Rückgang nicht signifikant. Wie bisherige Studien zeigen, ist die NMÜ in der Lage, typische Morphologie- und Grammatikschwierigkeiten zu bewältigen und darüber hinaus eine im Wesentlichen flüssige Übersetzung zu liefern (vgl. Bentivogli u. a. 2016; Toral & Sanchez-Cartagena 2017; Van Brussel u. a. 2018). Diese auffällige Flüssigkeit der Übersetzung zählt zu den Stärken des NMÜ-Ansatzes (vgl. Toral & Sanchez-Cartagena 2017). Angesichts dieser Entwicklung ist die Berücksichtigung der Stilqualität neben der Inhaltsqualität für eine entwicklungs-

[9] Man unterscheidet im Englischen zwischen restriktiven und nicht-restriktiven Relativsätzen: Im restriktiven Relativsatz wird die Bedeutung des Nomens, das beschrieben wird, begrenzt. Ohne den restriktiven Relativsatz ändert sich die Bedeutung des gesamten Satzes. Für restriktive Relativsätze verwendet man ‚that‘. Der nicht-restriktiver Relativsatz bietet lediglich zusätzliche Informationen über das Nomen und kann ohne Einfluss auf die Bedeutung entfernt werden. Für nicht-restriktive Relativsätze verwendet man ‚which‘. (vgl. Swan 1980: 527 ff.)

bzw. forschungsstandgemäße Bewertung erforderlich, obwohl die KS im Allgemeinen die Verständlichkeit und nicht den Stil im Fokus hat. Auf Basis der erzielten Ergebnisse bot das NMÜ-System sowohl mit als auch ohne Anwendung der Regeln nicht nur eine hohe Verständlichkeit und Genauigkeit, sondern auch einen dienlichen Stil. Im Folgenden eine Diskussion der Regeln, bei denen sich die Stilqualität signifikant und die Inhaltsqualität insignifikant bei dem NMÜ-System veränderten.

Regel 1 „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden" war die einzige Regel, bei der sich die Stilqualität bei dem NMÜ-System nach der Regelanwendung aufgrund der klaren orthografischen Darstellung der Oberflächentexte signifikant verbesserte. Gleichzeitig stieg die Inhaltsqualität insignifikant. Diese Verbesserung könnte jedoch auch erreicht werden, indem (1) die Oberflächentexte durch die Formatierung (z. B. fett, kursiv usw.) hervorgehoben und (2) firmen-, produktspezifische bzw. ungebräuchliche Termini der Oberflächentexte in das System integriert würden (siehe Beispiele unter §6.4 in der Analyse der Regel *„anz – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden"*). Gleichzeitig gab es drei andere Regeln, bei denen die Stilqualität signifikant und die Inhaltsqualität insignifikant abnahmen: Regel 4 „Eindeutige pronominale Bezüge verwenden", da die Wiederholung des Bezugsworts für die Bewerter stilistisch kritisch war; Regel 5 „Partizipialkonstruktionen vermeiden", da die Bewerter die Übersetzung der Partizipialkonstruktion (vor KS) für idiomatischer hielten; Regel 9 „Keine Wortteile weglassen", da die Wiederholung einiger Wortteile als unnatürlich wahrgenommen wurde. Zumal die Inhaltsqualität (Verständlichkeit und Genauigkeit) bei diesen drei Regeln vor der Regelanwendung höher ausfiel, deuten die Ergebnisse auf einen Fortschritt bei der Koreferenzauflösung (Regel 4) sowie bei der Übersetzung von Ellipsen (Regel 9) und von komplexen Strukturen wie der Partizipialkonstruktion (Regel 5) hin. Fortschritte bei der Übersetzung elliptischer Konstruktionen sowie bei der Koreferenzauflösung konnten zudem aktuelle Studien durch den Einsatz kontextfähiger NMÜ-Modelle und Strategien zur NMÜ auf Dokumentebene realisieren (vgl. Müller u. a. 2018; Stojanovski & Fraser 2018; Voita u. a. 2018; Matusov 2019; Stojanovski & Fraser 2019; Voita u. a. 2019). Von dieser Entwicklung kann die technische Dokumentation profitieren, denn eine Anwendung dieser Regeln zwecks der maschinellen Übersetzbarkeit ist im Falle der NMÜ nicht erforderlich. Unabhängig davon können diese Regeln angewendet werden, wenn die Einheitlichkeit bzw. Verständlichkeit Vorrang hat.

Die Veränderungen der AEM-Scores vor vs. nach KS waren bei allen Systemen minimal und entsprechend nicht signifikant. Gleichzeitig zeigte der Spearman-Korrelationstest in allen MÜ-Systemen hochsignifikante mittlere oder starke po-

sitive Korrelationen zwischen den Differenzen der Scores von TERbase und hLEPOR und den Qualitätsdifferenzen (Tabelle 5.178 unter §5.5.10). Diese positiven Korrelationen lassen die Ergebnisse der beiden Analysen (die Humanevaluation und die automatische Evaluation) sich gegenseitig bestätigen, denn eine Qualitätssteigerung ging mit einem verbesserten AEM-Score – und umgekehrt – einher.

6.6 Fazit

In diesem Kapitel wurden die Ergebnisse reflektierend zusammengefasst. In den Ergebnissen wurde eine kollektive positive Wirkung der Anwendung der analysierten KS-Regeln auf den MÜ-Output systemübergreifend festgestellt. Auf Regelebene systemübergreifend zeigten nur vier der neun untersuchten Regeln eine positive Wirkung auf den MÜ-Output. Drei Regeln zeigten hingegen einen negativen Effekt und die letzten zwei Regeln konnten keinen eindeutigen Effekt nachweisen. Betrachtet man die Regeln auf einer tieferen Ebene, nämlich auf Regel- und Systemebene, so waren die Ergebnisse unterschiedlich von einem MÜ-System zum anderen. Die Vergleichsanalyse der MÜ-Ansätze regelübergreifend am Beispiel der untersuchten Systeme zeigte, wie das NMÜ-System mit einer sehr niedrigen Fehleranzahl und sehr hohen Qualitätswerten sowohl vor als auch nach der Regelanwendung die Systeme der RBMÜ-, SMÜ- und HMÜ-Ansätze an Leistung deutlich übertraf. Während die KS-Anwendung mit einer Verbesserung der Inhalts- und Stilqualität bei den RBMÜ-, SMÜ- und HMÜ-Systemen auf unterschiedlichem Niveau einherging, fielen die Stil- und Inhaltsqualität (d. h. die Verständlichkeit und Genauigkeit) beim NMÜ-System vor der Anwendung der Regeln höher aus. Somit liefern die Ergebnisse der untersuchten Regeln ein Indiz dafür, dass eine Anwendung der KS zu Zwecken der maschinellen Übersetzbarkeit bei den früheren MÜ-Ansätzen (RBMÜ, SMÜ und HMÜ) jedoch nicht beim NMÜ-Ansatz förderlich sein kann. Je nach Anwendungskontext bleibt es dementsprechend kritisch zu bewerten, inwiefern eine solche KS-Anwendung erforderlich ist.

7 Fazit

It does not really matter what kind of ambiguity the system is up against; what matters is whether the system has the relevant data for disambiguation. (Hutchins & Somers 1992: 94)

7.1 Schlussfolgerungen

Die Studie beschäftigt sich mit der Analyse bzw. dem Vergleich des MÜ-Outputs unterschiedlicher MÜ-Ansätze (RBMÜ, SMÜ, HMÜ sowie NMÜ) vor und nach der Anwendung einzelner KS-Regeln hinsichtlich der aufgetretenen Fehler, Stil- und Inhaltsqualität sowie AEM-Scores. Der Vergleich fand auf vier Ebenen (Sprachenpaar-, Regel-, MÜ-System- sowie Regel- und MÜ-Systemebene) statt und zeigte Folgendes:

In Übereinstimmung mit mehreren vorherigen Studien (vgl. Nyberg & Mitamura 1996; Bernth 1999; Bernth & Gdaniec 2001: 208; Drugan 2013: 98; Drewer & Ziegler 2014: 196; Wittkowsky 2017: 92) zeigte die Anwendung der KS-Regeln auf Sprachenpaarebene (d. h. regel- und systemübergreifend) einen signifikanten positiven Einfluss auf den MÜ-Output im Sinne einer verringerten Fehleranzahl, einer erhöhten Stil- und Inhaltsqualität sowie verbesserter Scores zweier AEMs (TERbase und hLEPOR).

Eine genauere Analyse der Auswirkungen der einzelnen Regeln (systemübergreifend) ergab, dass nur die vier Regeln „Für zitierte Oberflächentexte gerade Anführungszeichen verwenden“, „Konstruktionen mit ‚sein + zu + Infinitiv‘ vermeiden“, „Konditionalsätze mit ‚Wenn‘ einleiten“ und „Funktionsverbgefüge vermeiden“ sich positiv auf den MÜ-Output auswirkten (Rückgang der Fehleranzahl und Verbesserung der Stil- und Inhaltsqualität sowie der AEM-Scores). Diese Regeln reduzierten die Ambiguität, vereinfachten die Satzstruktur und ermöglichten damit ein besseres Parsen, was zu einer genaueren, verständlicheren, idiomatischeren und aufmerksamkeitserregenden Übersetzung beitrug. Hingegen zeigten die drei Regeln „Passiv vermeiden“, „Partizipialkonstruktionen vermeiden“

und „Keine Wortteile weglassen“ einen negativen Einfluss auf den MÜ-Output (Anstieg der Fehleranzahl und Verschlechterung der Stil- und Inhaltsqualität sowie der AEM-Scores). Die signifikante Gemeinsamkeit bei allen drei Regeln war der Rückgang der Stilqualität nach der KS-Anwendung. Die Bewerter fanden das Passiv in einigen Fällen stilistisch adäquater, die Partizipialkonstruktion idiomatischer und die Formulierung mit Wortteilen zum Teil natürlicher. Das Ergebnis reflektiert, dass diese drei Regeln auf stilistischer Ebene nachteilig sein können; dies ist prinzipiell nachvollziehbar, da die KS die Verständlichkeit und nicht den Stil im Fokus hat. Gleichzeitig weist das Ergebnis auf einen gewissen Fortschritt bei der MÜ des Passivs, der Partizipialkonstruktion und der analysierten Form der Ellipsen hin, wodurch die Nicht-Anwendung der Regeln mit einer höheren Inhaltsqualität (signifikant im Falle der Regel „Passiv vermeiden“; nicht signifikant bei den anderen zwei Regeln) und somit einer höheren Genauigkeit bzw. Verständlichkeit – neben der besseren Stilqualität – verbunden war.

Für die letzten zwei Regeln „Eindeutige pronominale Bezüge verwenden“ und „Überflüssige Präfixe vermeiden“ wurden keine signifikanten Auswirkungen festgestellt. Bezüglich der Regel „Eindeutige pronominale Bezüge verwenden“ zeigt die qualitative Analyse der MÜ – in Übereinstimmung mit Bernth & Gdaniec (2001) –, dass die Entscheidung, ein Pronomen zu verwenden oder es durch seine Referenz zu ersetzen, auf einer Fall-zu-Fall-Basis getroffen wurde. So wurde eine Wiederholung der pronominalen Referenz empfohlen, nur wenn die Pronomen mehrdeutig waren, was die insignifikanten Ergebnisse dieser Regel begründet. Bei der Regel „Überflüssige Präfixe vermeiden“ waren die Ergebnisse insignifikant, da mehr als zwei Drittel der Sätze vor und nach der Regelanwendung korrekt und identisch übersetzt wurden, was zu einer vergleichbaren MÜ-Qualität führte. Trotz der insignifikanten Qualitätsveränderungen zeigen die Ergebnisse, dass die Systeme in den meisten Fällen (in 70 % der Fälle) die Koreferenz auflösen (Regel „Eindeutige pronominale Bezüge verwenden“) und die Verben mit getrennten Präfixen (in 71 % der Fälle) korrekt parsen (Regel „Überflüssige Präfixe vermeiden“) konnten, was auf einen Fortschritt bei zwei bekannten MÜ-Schwächen hindeutet.

Der Vergleich der Ergebnisse der früheren MÜ-Ansätze mit denen des NMÜ-Ansatzes ergab, dass die früheren MÜ-Systeme in vielen Fällen von den KS-Regeln zur Vermeidung unterschiedlicher MÜ-Fehler und zur Verbesserung ihrer Inhaltsqualität (Genauigkeit bzw. Verständlichkeit) sowie Stilqualität (Idiomatik der MÜ, Eignung der MÜ für die Intention des Inhalts bzw. orthografische Darstellung der MÜ) profitierten, während das NMÜ-System die meisten Sätze vor und nach der Anwendung aller Regeln fehlerfrei übersetzen konnte (in 83 % der Fälle), was mehr als doppelt so hoch wie bei jedem anderen System war. Darüber

hinaus verzeichnete das NMÜ-System in beiden Szenarien bei allen Regeln die höchsten Qualitätsratings unter allen Systemen in Bezug auf den Stil und den Inhalt (über 4,3 von 5 Punkten). Zudem war das NMÜ-System das einzige System, bei dem die Stil- und Inhaltsqualität nach der Regelanwendung zurückgingen – mit einem signifikanten Rückgang der Stilqualität. Der signifikante Rückgang der Stilqualität wurde auf Basis der Humanevaluation mit der höheren Idiomatik und Natürlichkeit der MÜ vor der Regelanwendung begründet. Demzufolge bietet das NMÜ-System die Möglichkeit, für gewünschte Textsorten auf die Anwendung der analysierten KS-Regeln zu verzichten und entsprechend von einem natürlichen Stil zu profitieren.

7.2 Rückblick und Ausblick

Vorherige Studien sahen in der MÜ eine wichtige Computeranwendung der KS. Die Kernidee ist, dass eine Abstimmung der KS-Regel auf das MÜ-System im Einsatz in einer effizienteren und effektiveren MÜ resultiert. Über diese Idee waren sich mehrere häufig zitierte Arbeiten über die letzten 20 Jahre einig. Dies formulierten Nyberg u. a. (2003: 254f.) folgendermaßen:

> MT is potentially one of the most interesting computational applications of CL. If a CL and an MT system are attuned to each other, MT of texts written in that CL can be much more efficient and effective, requiring far less – or ideally even no – human intervention. (Nyberg u. a. 2003: 254f.)

Als O'Brien (2003) acht KS-Regelsätze analysierte, mit dem Ziel anhand der gemeinsamen Regeln einen Kernregelsatz festzulegen, nach dem die Unternehmen arbeiten können, ohne das Rad neu erfinden zu müssen, fand sie die acht Regelsätze weitgehend individuell mit nur einer gemeinsamen Regel. Ihre Begründung dieses Ergebnisses beinhaltete:

> If source text is destined to be translated by a specific MT system for specific language pairs, then the rules will reflect the inherent weaknesses of the MT system and the known transfer problems between specific language pairs. (O'Brien 2003: 7)

Mit Fokus auf der Beziehung zwischen dem Pre-Editing und dem Post-Editing führt Göpferich (2007b: 481) das Argument fort – eine Unterstützung der MÜ-Systeme mithilfe von lexikalischen und syntaktischen Regeln reduziere den Post-Editing-Aufwand:

> Texte können nur dann mit geringem Nachbearbeitungsaufwand (Post-Editing) wirtschaftlich maschinell übersetzt werden, wenn der in ihnen verwendete Wortschatz (Lexik) und Satzbau (Syntax) sich im Rahmen dessen bewegt, was das maschinelle Übersetzungssystem analysieren kann. Bei der Abfassung maschinell zu übersetzender Texte sind also den Wortschatz und den Satzbau beschränkende Regeln zu beachten (Pre-Editing), was ebenfalls bereits als Standardisierung bezeichnet werden kann. (Göpferich 2007b: 481)

Bis 2017 ist die gleiche Schlussfolgerung zu finden. Sie wird von Wittkowsky (2017) in ähnlichen Worten vorgebracht:

> Damit Ausgangstexte dann auch maschinell gut übersetzbar sind, gilt es provokativ gesprochen eigentlich nur noch, die Regeln einzuhalten und auch das Übersetzungssystem speziell auf die Sprachverwendung abzustimmen. (Wittkowsky 2017: 93)

Die oben zitierten Studien bezogen sich auf Systeme der RBMÜ-, SMÜ- und HMÜ-Ansätze. Die Anwendung von KS-Regeln, die gezielt auf das verwendete System abgestimmt sind, war der Weg, der eingeschlagen wurde, um die Systemschwächen auszugleichen und auf diese Weise die maschinelle Übersetzbarkeit erhöhen zu können. In der vorliegenden Studie wurden die fünf Systeme im Ist-Zustand, d. h. ohne Abstimmung zwischen den analysierten Regeln und den Systemen und ohne vorheriges Training mit Korpora, die auf das Testmaterial abgestimmt sind, untersucht.

Die Untersuchung der Auswirkungen der einzelnen Regeln auf MÜ-Systemebene ergab, dass diese Auswirkungen bei den früheren MÜ-Ansätzen (RBMÜ-, SMÜ- und HMÜ-Systemen) von einem Ansatz zum anderen je nach Stärken und Schwächen des jeweiligen Systems unterschiedlich – signifikant – positiv und negativ waren. Es bleibt daher die Notwendigkeit bestehen, die wirksamen Regeln in jedem Implementierungskontext (Sprachenpaar, Übersetzungsrichtung, Domäne und MÜ-Ansatz bzw. -System) zu identifizieren und sie gezielt auf diesen Kontext abzustimmen.

Eine Restriktion durch die analysierten KS-Regeln konnte hingegen mithilfe des untersuchten NMÜ-Systems vermieden werden, denn das NMÜ-System Google Translate war in der Lage, die meisten Sätze vor und nach der Anwendung aller Regeln fehlerfrei und mit den höchsten Qualitätswerten unter allen Systemen – ohne vorheriges Training mit untersuchungsrelevanten Korpora oder vorherige Abstimmung auf die analysierten Regeln – zu übersetzen. Überdies ging die Inhaltsqualität leicht und die Stilqualität signifikant nach der Anwendung der analysierten Regeln zurück, was nach der Humanevaluation auf die

höhere Idiomatik und Natürlichkeit der MÜ vor KS zurückgeführt wurde. Außerdem nahm die Fehleranzahl nach der KS-Anwendung insignifikant zu.

Die Analyse im Rahmen dieser Studie erfolgte auf Satzebene, somit beschränkt sie sich auf satzrelevante KS-Regeln. Eine Analyse in einem größeren Rahmen ist ausdrücklich erwünscht, um den Einfluss von kontextrelevanten KS-Regeln (d. h. Regeln, die mehrere Sätze betreffen) auf den MÜ-Output zu untersuchen. Die kontextbezogene MÜ bzw. die MÜ auf Dokumentebene zählt zu den bekannten herausfordernden Zielen der MÜ (Zhang & Zong 2020). Inwiefern die KS eine kontextbezogene MÜ unterstützen kann, ist eine Frage, die sich empirisch beantworten lässt. Nach dem aktuellen Stand der NMÜ-Forschung wurden bereits im Bereich der kontextbezogenen MÜ bzw. der MÜ auf Dokumentebene Fortschritte realisiert, und das sogar in der Literaturübersetzung; einer Domäne, die als besonders herausfordernd für die MÜ gilt (vgl. Toral & Way 2018; Matusov 2019). Ferner entwickelten mehrere Studien kontextfähige NMÜ-Modelle sowie Strategien zur NMÜ auf Dokumentebene, mit denen klassische MÜ-Schwierigkeiten, wie Deixis, Ellipsen, Koreferenzauflösung, Kohärenz und lexikalische Kohäsion, bewältigt werden konnten (Müller u. a. 2018; Stojanovski & Fraser 2018; Voita u. a. 2018; Stojanovski & Fraser 2019; Voita u. a. 2019). Auf Basis der Ergebnisse der vorliegenden Studie sowie der geschilderten rapiden Entwicklungsfortschritte der NMÜ (siehe auch §3.3.4) ist es zu erwarten, dass eine Anwendung der KS zum Zwecke der maschinellen Übersetzbarkeit in naher Zukunft zunehmend in den Hintergrund gedrängt wird. Sollte dies der Fall sein, können die Unternehmen die KS gezielt für die weiteren Zwecke – wie die Konsistenz, Lesbarkeit von umfangreichen Dokumenten und Verständlichkeit von komplexen Inhalten – anwenden und sich auf die tatsächlich dafür erforderlichen Regeln beschränken.

Wenn der Einsatz umfangreicher KS-Regelsätze mit verringerter Regel-Usability und Autorenproduktivität (Mitamura 1999), übermäßigem Intervenieren und Erschwerung des Schreibprozesses (O'Brien & Roturier 2007), Zeit- und Kostenaufwand – trotz der Verwendung von CL-Checkern (Govyaerts 1996; O'Brien & Roturier 2007), verminderter Textakzeptabilität (Roturier 2006) und unnatürlichem Stil (Lehrndorfer & Reuther 2008: 112f.) verbunden sein kann, ist es ersichtlich, dass *eine Reduzierung des KS-Einsatzes auf das Wesentliche folgende Implikationen hätte:*

Zeit- und Kostenersparnisse: In der Praxis übersetzen die Unternehmen ihre Dokumentation in der Regel in mehrere Sprachen. Um den verschiedenen Transferproblemen entgegenwirken zu können, müssen für Zwecke der maschinellen Übersetzbarkeit für jedes Sprachenpaar die adäquaten KS-Regeln angewendet werden (vgl. O'Brien 2003). Sollte aber ein robustes NMÜ-System dem Unternehmen eine qualitative Übersetzung in verschiedene Sprachen unter Anwendung

von (deutlich) weniger KS-Regeln ermöglichen, würde dies mit Zeit- und Kostenersparnissen einhergehen.

Vereinfachung des Schreibprozesses und ggf. Schaffung eines Raums für Kreativität: Die Einhaltung der KS-Regeln erfordert in manchen Fällen eine vollständige Umformulierung der Sätze. Aus Sicht der Autoren kann die Einhaltung der KS-Regeln komplex, kreativitätshemmend und zeitaufwendig ausfallen. In der Praxis wird außerdem die technische Dokumentation von technischen Redakteuren und nicht selten auch von den Fachabteilungen bzw. Fachexperten durchgeführt, die über begrenztes linguistisches Wissen zu dem Verstehen und der Umsetzung aller Regeln verfügen (Van der Eijk u. a. 1996; Aranberri & Roturier 2009). Für beide Gruppen ist die Einschränkung der angewendeten Regeln von Vorteil, denn sie würde zu einer Erhöhung der Autorenproduktivität beitragen. Technische Redakteure, die sich durch die Regeleinhaltung in ihrer Kreativität gebremst fühlen (vgl. Nyberg u. a. 2003: 249; Lehrndorfer & Reuther 2008: 112; Drewer & Ziegler 2014: 209), hätten ggf. mehr Raum für Kreativität. Diesen Raum kann das Unternehmen anhand des CL-Checkers je nach Textsorte, Natur des verfassten Dokuments und Autor bzw. Abteilung steuern. Beide Effekte (Vereinfachung des Schreibprozesses und Schaffung eines Kreativitätsraums) würden im Endeffekt zur Erhöhung der Autorenmotivation beitragen.

Natürlicherer Stil und erhöhte Textakzeptabilität: Die gesteigerte Motivation und Kreativität des Autors würden sich in einem natürlichen Stil widerspiegeln. Die Konsistenz ist zwar herkömmlich in der technischen Dokumentation, gleichzeitig empfiehlt Püschel (1996: 335f.) den technischen Redakteuren, „den Text so abwechslungsreich wie möglich zu machen", denn „auch ein Stilbruch kann die Aufmerksamkeit wecken". Wie die Ergebnisse des Einflusses der KS auf den NMÜ-Outputs zeigten, wurden MÜ, die vor und nach der KS-Anwendung fehlerfrei waren, vor der KS-Anwendung als natürlicher bewertet, was wiederum die Akzeptanz des Texts durch die Rezipienten erhöhen würde. Je nach Textsorte und Dokumentationsziel kann ein natürlicher Stil erwünscht sein, z. B. für technische Dokumentationen, die Produkte beschreiben.

Gleichzeitig ist Folgendes zu berücksichtigen: Da die deutsche KS primär die Verständlichkeit des Ausgangstexts und nicht seinen Stil zum Ziel hat, steht das Unternehmen bei der Entscheidung über die Anwendung bzw. Nicht-Anwendung einer Regel angesichts der erzielten Ergebnisse vor einem gewissen Tradeoff zwischen der Maximierung der Verständlichkeit des Ausgangstexts und der Schaffung von Raum für natürlichen Stil. Diese Entscheidung lässt sich fallspezifisch je nach Textsorte und -ziel treffen. Die NMÜ ebnet den Weg für einen natürlichen Stil kombiniert mit einer hohen Genauigkeit und Verständlichkeit

des Zieltexts. Es bleibt dem Unternehmen überlassen, wann es diesen Weg geht und davon profitiert.

Die genannten potenziellen Vorteile der Einschränkung des KS-Einsatzes auf für vordefinierte Ziele festgelegte Regeln sowie die erzielten Ergebnisse des NMÜ-Systems hätten für den technischen Redakteur, den Übersetzer sowie für die Ausbildung von beiden folgende Implikationen:

Je nach Unternehmen bzw. Verantwortungsaufteilung spielen die *technischen Redakteure* unterschiedliche Rollen bei dem Pre-Editing bzw. Einsatz von KS-Regeln. Dazu zählen die Definition der Ziele des KS-Einsatzes, die Auswahl der dafür erforderlichen Regeln sowie die Durchführung des Pre-Editing. Darüber hinaus trifft der technische Redakteur die Entscheidung über Texte, die sich für MÜ – mit einem Kostensparpotenzial – eignen. Für diese Texte kann er wiederum die Textsorten differenziert behandeln und KS-Regeln, die dem Ziel des Texts dienlich sind, anwenden. Haben die Verständlichkeit bzw. die Lesbarkeit Priorität, können die entsprechenden Regeln angewendet werden. Soll der Stil priorisiert werden, zeigen die Ergebnisse der analysierten Regeln die Tendenz zu einer möglichen Deregulierung, die mit einer hohen Stil- und Inhaltsqualität der MÜ einhergeht.

Bei dem *Übersetzer* kann ebenfalls zwischen Textsorten bzw. Dokumentationsarten, die für eine MÜ geeignet und anderen, die nicht geeignet sind, unterschieden werden. Bei Ersterem bleibt die MÜ als ein Tool, das nur von qualifizierten Übersetzern sinnvoll eingesetzt werden kann. Daher muss der Fortschritt dieses Tools mit einer Weiterentwicklung der Übersetzerqualifikationen einhergehen. Für die zurzeit bekannten Schwächen der NMÜ-Systeme, dass sie in manchen Fällen zwar sehr flüssigen aber ungenauen bzw. unvollständigen Output liefern, ist ein geschultes Auge gefragt. Die Post-Editing-Aufgabe muss daher tiefgehend durchgeführt werden, um diese Art von Genauigkeitsfehlern aufzudecken (vgl. Volk 2018). Gleichzeitig darf nicht unerwähnt bleiben, dass die NMÜ-Systeme von den übersetzten Texten lernen und sich weiter verbessern. Der Verbesserungsgrad wird jedoch von einer Domäne zur anderen bzw. von einem Sprachenpaar zum anderen und somit der Umfang des Post-Editing variieren. Unerlässlich bleibt die Rolle der Übersetzer bei Texten, die weniger für die MÜ geeignet sind, nämlich Texte, die Kreativität erfordern bzw. meist nicht standardisiert sein können (z. B. Marketing-Texte oder Kundenkorrespondenz). Für die Weiterentwicklung der NMÜ wird außerdem der Humanübersetzer benötigt: Nicht für alle Sprachenpaare existieren ausreichende bzw. qualitative Trainingsdaten. Das gilt nicht nur für seltene Sprachen, sondern sogar für Sprachen, die von vielen Völkern gesprochen werden, wie z. B. Arabisch und Russisch. Für

die Erstellung dieser bilingualen Trainingsdaten sind qualitative Humanübersetzungen erforderlich.[1] Zudem werden die Fachkenntnisse und Kompetenzen der Übersetzer vermehrt zur Unterstützung von MÜ-Entwicklern bei der laufenden Optimierung der Systeme sowie von den MÜ-Forschern bei der Evaluation des MÜ-Outputs gebraucht. In der *Industrie* werden die Unternehmen zunehmend ihre Übersetzungsprozesse und -workflows optimieren. Auch hierfür ist das translationstechnologische Wissen der Übersetzer bei der Gestaltung der Workflows sowie bei der Integration der Terminologiearbeit hilfreich und ausdrücklich erwünscht.

Für die *Didaktik* lenken die Ergebnisse die Aufmerksamkeit auf die Notwendigkeit der Weiterentwicklung der Bildungsinhalte, um mit dem gegenwärtigen Fortschritt in den Translationstechnologien und seiner Auswirkung auf die technische Dokumentation sowie den Übersetzungsprozess mithalten zu können. Mit dem Fortschritt der NMÜ steht fest, dass die NMÜ-Systeme durch das stetig wachsende Textvolumen weiter lernen und sich verbessern werden. Es liegt daher auf der Hand, dass der Übersetzungsmarkt zunehmend von der NMÜ dominiert wird. Für diese Entwicklung müssen die technischen Redakteure und Übersetzer ausgerüstet werden. Bei beiden Zielgruppen sind neben den klassischen Aufgabenfeldern die Gestaltung, der Umgang und die Unterstützung automatisierter Übersetzungsprozesse zunehmend wesentlich für ihren Beruf. So gehören redaktions- und translationstechnisches Fachwissen sowie redaktions- und translationstechnologische Kenntnisse und Fähigkeiten neben dem Basiswissen in der technischen Redaktion und Translation zu einem marktadäquaten bzw. realitätsabbildenden Curriculum. Konkret sind Lerninhalte zu den folgenden Themenbereichen bzw. Fragestellungen zu empfehlen: Wie der Übersetzungsprozess effizient konzipiert werden kann; in welchen Szenarien Pre- bzw. Post-Editing erforderlich ist; was mit dem Pre- und Post-Editing in Zusammenhang mit der NMÜ zu beachten ist; welches Maß an Sprachkontrolle notwendig ist; in welchen Fällen die Sprachkontrolle überflüssig sein kann; wie das maschinell übersetzungsgerechte Schreiben von der Entwicklung der NMÜ beeinflusst wird; wie NMÜ-Systeme effektiv in den Unternehmensworkflow integriert werden können; wie die NMÜ-Systeme lernen bzw. trainieren und sich verbessern können. Des Weiteren bleibt die Rolle des Terminologiemanagements wesentlich für die (N)MÜ. Wie die Ergebnisse zeigten, war es die Terminologie, die dem analysierten generischen NMÜ-System in den meisten Fällen fehlte, um eine fehlerfreie MÜ zu

[1] Wohlgemerkt, läuft die Weiterentwicklung der sog. „semi-supervised" bzw. „unsupervised" NMÜ-Modelle, die (überwiegend) mit massiven monolingualen Trainingsdaten arbeiten, auf Hochtouren (vgl. Zhang & Zong 2020).

produzieren. Daher gehören das Terminologiemanagement und seine Integration in den Übersetzungsworkflow zu einem praxisorientierten Curriculum.

7.3 Beitrag und Einschränkungen der Studie sowie zukünftige Forschung

Die Studie untersucht KS-Regeln der deutschen Sprache – einer Sprache, bei der ein Mangel an empirischen Untersuchungen im Bereich der KS vorliegt. Dabei werden die Auswirkungen der einzelnen Regeln auf den MÜ-Output verschiedener MÜ-Ansätze (RBMÜ, SMÜ, HMÜ und NMÜ) analysiert. Somit deckt die Studie den jüngsten MÜ-Ansatz (NMÜ) ab und vergleicht ihn mit früheren Ansätzen im Kontext der KS. Die vorherigen Studien zu den Auswirkungen der KS auf die MÜ-Qualität beziehen sich auf die früheren MÜ-Ansätze (RBMÜ, SMÜ und HMÜ). Der aktuellste Ansatz der NMÜ wurde nach bestem Wissen noch nicht im Kontext der KS untersucht. Angesichts der hervorragenden Qualität des MÜ-Outputs neuronaler Systeme (vgl. Bentivogli u. a. 2016; Wu u. a. 2016; Castilho, Moorkens, Gaspari, Sennrich u. a. 2017; Toral & Sanchez-Cartagena 2017; Popović 2018), ist es an der Zeit, dass die KS-Community die Auswirkung der KS-Regeln auf die MÜ wieder aufgreift und überprüft. Dies realisierte die Studie anhand eines Mixed-Methods-Triangulationsansatzes auf Basis von drei bewährten Evaluationsmethoden: Fehlerannotation, Humanevaluation und automatischer Evaluation. Die systematische Dokumentation aller Details der methodischen Vorgehensweise der einzelnen Analysen zusammen mit der Begründung sämtlicher getroffenen Entscheidungen ermöglichen eine Replikation der Studie. Anders als der bisher oft belegte allgemeine positive Effekt der KS auf die MÜ zeigt die Studie, dass nicht alle analysierten KS-Regeln zur Verbesserung der MÜ erforderlich sind. Bei den früheren MÜ-Ansätzen sind die Regeln reduzierbar und bei dem NMÜ-Ansatz überflüssig und wirken sich negativ auf den Stil aus.

Das *Design der Humanevaluation* stellt einen weiteren Beitrag der vorliegenden Studie dar. Im Testdesign wurden die Qualitätsdefinitionen in die Qualitätskriterien integriert (Abschnitt [3a] und [3b] in Abbildung 4.6 unter §4.5.5.2 „Darstellung der Evaluation"), d. h. nicht nur in Form von Definitionen am Anfang des Tests zur Verfügung gestellt, wie es in Evaluationsstudien typisch ist. Somit hatten alle Teilnehmer eine direkte und einheitliche Basis für die Vergabe der Qualität-Scores, die sicherlich zu den hohen Intrarater- und Interrater-Agreements beitrug. Das Testdesign sieht eine methodeninterne Triangulation vor. Die Teilnehmer wurden aufgefordert, die zutreffenden Qualitätskriterien

anzukreuzen (Abschnitt [3a] und [3b] in Abbildung 4.6), zu kommentieren (Abschnitt [4]) bzw. posteditieren (Abschnitt [5]) und anschließend einen Score zu vergeben (Abschnitt [2]). Diese methodeninterne Triangulation fördert die interne Konsistenz, die Zuverlässigkeit und eine genaue Interpretation der Daten. Ferner ist das Testdesign replizierbar.

Die folgenden Einschränkungen sind jedoch zu erwähnen:

Da der Fokus überwiegend auf dem Vergleich der KS-Auswirkung auf den Systemoutput der unterschiedlichen MÜ-Ansätze und insbesondere auf einer Gegenüberstellung mit dem jüngsten Ansatz der NMÜ lag, wurden fünf MÜ-Systeme untersucht und dabei die Variablen Sprachenpaar, Übersetzungsrichtung und Domäne für eine einheitliche Vergleichsbasis *konstant* gehalten. Mit diesen Dimensionen und anhand eines angemessen großen Datensatzes stand zudem der Studienumfang mit den zur Verfügung stehenden Zeit- und Finanzressourcen im Verhältnis.

Es wurde nur *ein Sprachenpaar und eine Übersetzungsrichtung* untersucht, dabei lag der Fokus auf der deutschen Sprache (als Ausgangssprache), die für international agierende Unternehmen aus deutschsprachigen Ländern von großer Relevanz ist. Für diese Unternehmen stellt Englisch eine bedeutsame Zielsprache dar, in der ein großes Dokumentationsvolumen für englischsprachige EU-Länder gemäß der Maschinenrichtlinie 2006/42/EG sowie für weitere große internationale englischsprachige Märkte zur Verfügung gestellt werden müssen.

Da die *technische Domäne* den Hauptanwendungsbereich für die KS darstellt und die analysierten tekom-Regeln für die technische Dokumentation bestimmt sind, wurden nur technische Texte in der Studie analysiert. Der gebildete Korpus umfasst jedoch zehn technische Dokumente für verschiedene Produkte unterschiedlicher Hersteller.

Aus zwei Gründen wurden nur *neun KS-Regeln* analysiert: Erstens, es wurden nur die Regeln, die bestimmte Kriterien zur Untersuchung im Rahmen einer Black-Box-Analyse erfüllten (siehe §4.5.2.1), ausgewählt. Zweitens, es gab weitere Regeln, die die festgelegten Kriterien erfüllten, allerdings beinhaltete der Korpus nicht genug Verstöße gegen diese Regeln. Für eine ausgewogene quantitative und qualitative Analyse legte die Forscherin Wert darauf, die gleiche Anzahl an Sätzen bei allen Regeln zu untersuchen. Bei dem Einsatz der KS-Regeln wurde nur ein *Umsetzungsmuster* pro Regel angewandt, um die Anzahl der unabhängigen Variablen im Rahmen zu halten. Dieser Rahmen ermöglichte in der Humanevaluation die komplette Bewertung von 1.100 Sätzen durch jeden Teilnehmer. Eine Erweiterung der Anzahl der Regeln oder der Umsetzungsmuster würde eine Erhöhung der Anzahl der Teilnehmer und eine Aufteilung der Sätze auf mehrere Teilnehmer erfordern. Der Forscherin war es jedoch ein Anliegen,

dass die 1.100 Sätze komplett von allen Teilnehmern gleichermaßen bewertet werden, um einem potenziellen negativen Einfluss auf das Agreement vorzubeugen.

Die *Anzahl der Ausgangssätze* war zwar nicht hoch, jedoch wurden sie von fünf verschiedenen MÜ-Systemen übersetzt. Somit war letztendlich die Anzahl der analysierten MÜ-Sätze ziemlich hoch und diese wurden von acht Teilnehmern bei der Humanevaluation bewertet. Aus Realisierbarkeitsgründen und wie das Feedback der Teilnehmer zeigte, wäre es schwierig gewesen, Teilnehmer zu finden, die bereit gewesen wären, mehr (als 1.100) Sätze zu bewerten. Schließlich wird in MÜ-Evaluationsstudien wie von Fiederer und O'Brien (2010a: 59) darauf hingewiesen, wie wichtig es ist, einen Kompromiss zwischen der Größe des Datensatzes und der Integrität der Ergebnisse zu finden; so begrenzten Fiederer und O'Brien die Anzahl der bewerteten Sätze pro Teilnehmer (auf 180 Sätze), um das „risk of boredom and its negative consequences" zu vermeiden (ebd.).

Eine potenzielle Schwäche der Studie könnte in der *Aufbereitung der Testsuite* liegen. Die vorgenommenen Aufbereitungsschritte waren für eine Black-Box-Analyse zur Untersuchung des Effekts jeder einzelnen Regel unerlässlich. Die Aufbereitung erfolgte nach klar definierten Kriterien, gefolgt von Schritten zur Qualitätsprüfung (siehe §4.5.3), mit dem Ziel sicherzustellen, dass die Sätze so natürlich wie möglich bleiben und nicht gezielt vereinfacht werden (wie es an der Fehleranzahl erkennbar ist). Eine Glas-Box-Analyse würde eine wesentliche Reduzierung der Aufbereitungsschritte ermöglichen, jedoch wäre sie mit den online zugänglichen untersuchten Systemen schwer realisierbar. Stattdessen wurde das Ziel der Studie mithilfe der Definition der KS-Stelle erreicht (siehe §4.5.2.1). Letztendlich lag der Fokus der Analyse primär darauf, *ob* Fehler innerhalb der KS-Stelle auftraten bzw. behoben wurden und *wie* die KS-Stelle qualitativ zu bewerten war; und nicht daran, *warum* genau der Fehler auftrat bzw. behoben wurde.

Die Forscherin testete gezielt *generische MÜ-Systeme und ohne vorheriges Training mit Korpora, die auf das Testmaterial abgestimmt sind*, um die Kapazitäten der Systeme im Ist-Zustand zu untersuchen. Hätte die Studie die Systeme vor der Untersuchung trainiert, wären die Ergebnisse von den Trainingsdaten abhängig (d. h. bessere Ergebnisse beim Kontrollierten Szenario, wenn die Korpora kontrolliert sind, und umgekehrt). Außerdem wäre ein angemessener Vergleich der Ergebnisse der verschiedenen Systeme nicht realisierbar, denn die Systeme, die basierend auf Trainingskorpora arbeiten, hätten je nach Kontrollgrad einen Vorteil bzw. Nachteil gegenüber dem regelbasierten System.

Dementsprechend sind folgende Forschungsaspekte erstrebenswert:

Es wäre von großem Interesse, die Studie für weitere Sprachenpaare und Textsorten sowie mehr KS-Regeln mit unterschiedlichen Umsetzungsmustern zu replizieren.

Bei der Untersuchung weiterer Sprachenpaare ist ferner der Grad der „*Configurationality*“ jeder Sprache, d. h. der Grad der Flexibilität bei der Wortstellung, zu berücksichtigen. In dieser Studie wurde das Sprachenpaar Deutsch > Englisch unter die Lupe genommen. Englisch – im Gegensatz zum Deutschen – zeichnet sich durch eine ziemlich starre Wortstellung aus (vgl. Hawkins 1986: 41f.); Deutsch weist hingegen eine relativ hohe Wortstellungsfreiheit auf und ist entsprechend komplexer zu kontrollieren. Der Grad der „Configurationality“ kann eine wesentliche Rolle bei der Analyse der KS-Wirkung spielen, daher ist sein Einfluss untersuchenswert.

Zur Zeit der Durchführung der empirischen Studie dieser Arbeit (Ende 2016 – Anfang 2017) stand das NMÜ-System Google Translate als erstes NMÜ-System an der Spitze. In der Zwischenzeit wurden weitere NMÜ-Systeme entwickelt (wie z. B. das deutsche NMÜ-System DeepL)[2] sowie ältere Systeme zu einem NMÜ-System weiterentwickelt, darunter Systran und SDL.[3] Der NMÜ-Ansatz und seine Hybridisation mit anderen Ansätzen stellen zweifellos über die kommenden Jahre die Zukunft der MÜ dar. Vor diesem Hintergrund wäre es wertvoll, die Notwendigkeit der KS-Anwendung zu Zwecken der maschinellen Übersetzbarkeit bei mehreren NMÜ-Systemen zu untersuchen. Sollte der NMÜ-Output von mehreren Systemen bei diversen KS-Regeln vor sowie nach dem KS-Einsatz qualitativ vergleichbar sein, würde dies das Ende des KS-Einsatzes für Zwecke der maschinellen Übersetzbarkeit ankündigen.

Die Durchführung einer *Glas-Box-Analyse* (anstatt einer Black-Box-Analyse) ist eine weitere erstrebenswerte Idee für zukünftige Forschung. In der Glas-Box-Analyse liegen die Systemfunktionsweise und -abläufe offen. Dies würde eine Analyse mithilfe eines natürlichen Korpus ermöglichen und somit eine genauere Verfolgung und Erklärung des Einflusses der Regeln auf die MÜ ohne bzw. mit deutlich weniger *Datenaufbereitung*.

Vor dem Hintergrund, dass die KS eine Komponente des *übersetzungsgerechten Schreibens* darstellt, bedarf der Bereich des übersetzungsgerechten Schreibens angesichts der erzielten Ergebnisse eines Umdenkens bzw. einer Revision. Hierbei sind empirische Untersuchungen der unterschiedlichen Regeln des maschinell übersetzungsgerechten Schreibens und ihres aktuellen Effekts auf den NMÜ-Output notwendig.

[2] Das NMÜ-System DeepL findet sich unter: https://www.deepl.com/translator

[3] Die neuronalen Systeme von Systran, SDL und Bing kamen 2017 bzw. 2018 auf den Markt: http://www.systransoft.com/systran/translation-technology/pure-neural-machine-translation; https://www.sdl.com/de/about/news-media/press/2018/sdls-neural-machine-translation-sets-new-industry-standards-with-state-of-the-art-dictionary-and-image-translation-features.html; https://www.microsoft.com/de-de/translator/blog/2018/11/14/nextgennmt [abgerufen am 16.04.2019]

7.4 Schlusswort: Untersuchung der Sprachkontrolle im Spiegel der MÜ

Eine Gegenüberstellung der Ergebnisse der älteren Systeme und der des NMÜ-Systems bestätigt die Aussage von Hutchins & Somers (1992: 94): „It does not really matter what kind of ambiguity the system is up against; what matters is whether the system has the relevant data for disambiguation." Das NMÜ-System zeigte im Gegensatz zu den älteren Systemen, dass es über die relevanten Daten bzw. die adäquate Technik zu einer erfolgreichen Disambiguierung unabhängig von der Anwendung bzw. Nicht-Anwendung der KS-Regeln verfügt. Verwendet das Unternehmen ein vergleichbares System der früheren Ansätze, kann es von der KS profitieren, (insbesondere) wenn die Regeln speziell auf die Schwächen des sich im Einsatz befindenden Systems abgestimmt sind. Vergleichbare NMÜ-Systeme in Kombination mit Terminologieintegration bzw. -management ohne Anwendung der untersuchten Regeln scheinen hingegen vielversprechend im Sinne einer hohen Stil- und Inhaltsqualität der MÜ zu sein. Die Anwendung der analysierten Regeln zum Zwecke der maschinellen Übersetzbarkeit zeigte sich als nicht erforderlich. Ferner lässt sich angesichts des sprunghaften Fortschritts im Bereich der NMÜ, der sich mittlerweile auf die kontextfähige MÜ bzw. die MÜ auf Dokumentebene erstreckt, antizipieren, dass eine KS-Anwendung zum Zwecke der maschinellen Übersetzbarkeit in naher Zukunft zunehmend in den Hintergrund gedrängt wird.

Zusammenfassend bleibt die zentrale Rolle der KS bei der Etablierung der Unternehmenssprache (Corporate Language) als einem Bestandteil der Unternehmensidentität (Corporate Identity) sowie bei der Standardisierung der technischen Dokumentation in der Ausgangssprache zum Zwecke der Lesbarkeit, Verständlichkeit und Wiederverwendbarkeit unberührt. Sofern die maschinelle Übersetzbarkeit ein Ziel ist, kann aufgrund des MÜ-Fortschritts für gewünschte Textsorten auf die KS verzichtet werden und weiterhin eine hohe – oder sogar höhere – Stil- und Inhaltsqualität erzielt werden.

Anhang A: Testanweisungen der Humanevaluation

Thanks a lot for your interest in taking part in this study

Before you begin, the following short description provides you with the necessary information on what to expect and what to keep in mind:

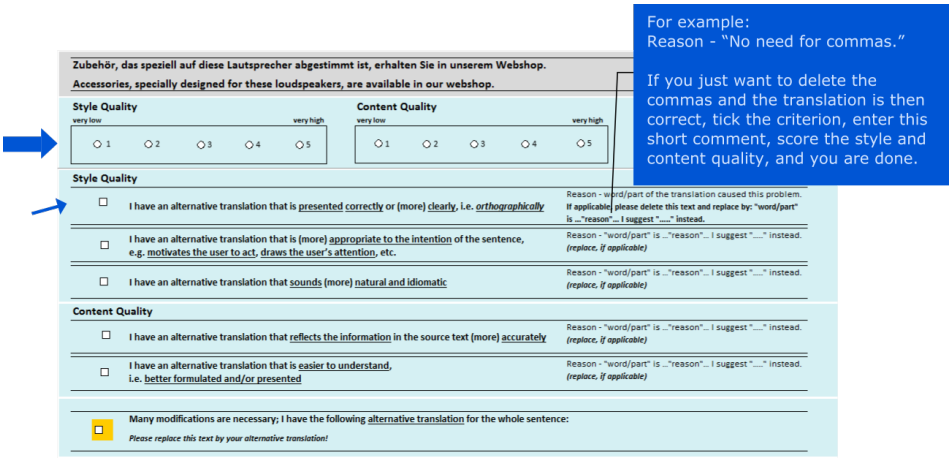
- You are going to rate translations of sentences extracted from **user manuals for different products**.
  - Some sentences seem identical at the first glance, but they are not, so please read each sentence **c a r e f u l l y**.
  - If you found a sentence, for which you would prefer to have more information about its **context**, please enter that as a comment and rate the translation as best as you can.
- Your main task is to **rate** the content and style quality on a scale from 1 (very low quality) to 5 (very high quality) **based on the criteria mentioned in the following definitions**:

  Style quality: The extent to which the translation sounds natural and idiomatic in Standard Written English, is appropriate to the intention of its content as well as is presented clearly orthographically.

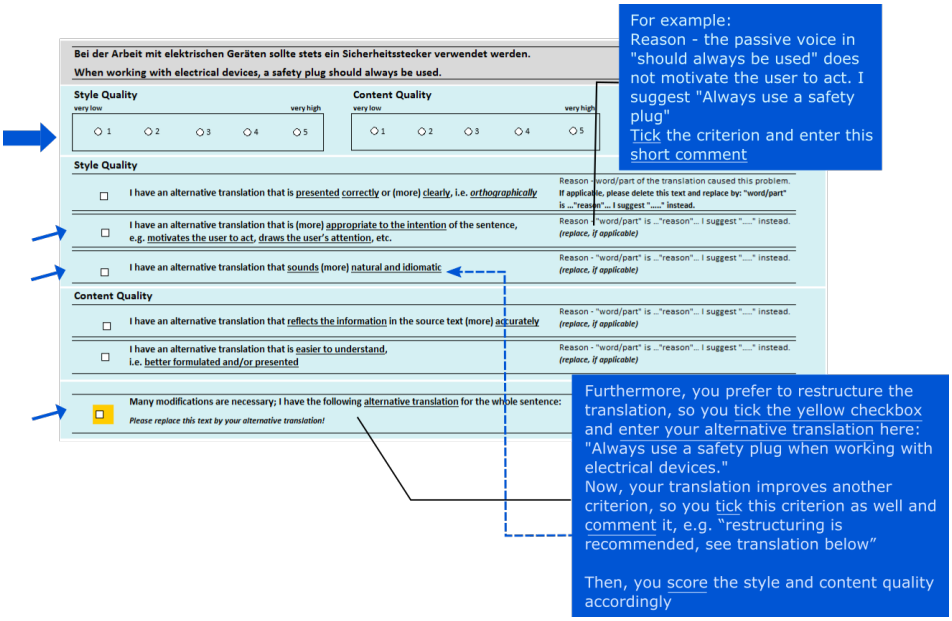
  Content quality: The extent to which the translation reflects the information in the source text accurately; and the extent to which the translation is easy to understand.
- Please work as quickly as comfortably possible. For statistical purposes, you need to enter the starting and ending time of each test on the first and last page, respectively.

HOW TO RATE

Example 1

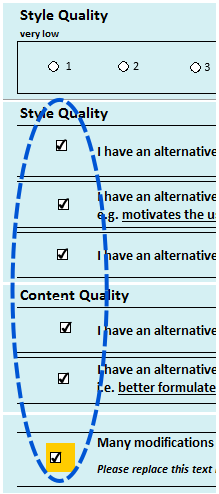


Example 2



Note

Technically, the checkboxes can be all ticked at the same time, allowing you to select all relevant criteria.



Each sentence in the test is provided separately in a table / tab in Excel below. Please scroll to the right to see all 25 tabs.

How to proceed?

Please proceed as follows:

1. Tick the quality criteria that are **not** satisfied, if any.
2. For the selected quality criteria, you need to enter a reason. This reason should include the following information: "wrong/problematic part" + "reason" + "your correction/suggestion for this part" *(just a short comment)*
3. Please score the "Style Quality" and "Content Quality" **after** selecting the relevant quality criteria and commenting them.
4. When is it necessary to enter an alternative translation for the whole sentence at the bottom?
    - If you just recommend replacing a certain part in the translation and agree with the rest of the translation, you do not need to enter an alternative translation.

- If your suggested modification (e.g. in the main clause) requires a further modification (e.g. in the subordinate clause), an alternative translation for the whole sentence is needed; otherwise, the translation might not sound idiomatic.

5. It is possible that a certain part of the translation causes *a number of different problems (e.g. 2 style problems and 1 content problem)*, reducing *both style* and *content quality*. In other words, you may suggest a modification (e.g. an orthographic issue), which makes the translation *presented more clearly* and at the same time *eases understanding its content.*

6. You may find a translation *correct* with regard to the *content*. However, *stylistically*, it still needs to be optimized to make the translation sound *natural* for an English native speaker and/or *suitable* for use in an English user manual.

7. If the error is a *missing word*, please tick the relevant quality criteria and enter as a reason e.g. the word "XYZ" is missing after the word "ABC".

Anhang B: Pre- und Posttests der Humanevaluation

Pretest

1. Geschlecht:
2. Herkunftsland (In welchem Land bist Du aufgewachsen?)
3. Deutschkenntnisse:

   Wie hast Du die deutsche Sprache gelernt? *Mehrere Antworten sind möglich:*

   a) Ich bin zweisprachig aufgewachsen.
   b) Durch Sprachkurse (wie lange war die Dauer der Kurse?)
   c) In der Schule
   d) Durch meinen Aufenthalt in Deutschland (wie lange war Dein Aufenthalt?)
   e) Sonstiges: .....
4. Berufliche Übersetzungserfahrung: Hast Du berufliche Übersetzungserfahrung gesammelt? Wenn ja, wie lange und überwiegend in welchen Bereichen?

Posttest

Teil 1: Fachliche Fragen

1. **Verwendung von maschinellen Übersetzungssystemen (MÜ-Systemen): Wie gehst Du in der Regel vor, wenn Du einen technischen Text übersetzen möchtest? bzw. wie würdest Du vorgehen?**

   a) Ich übersetze den Text zuerst mit einem MÜ-System und post-editiere die Übersetzung.

b) Ich benutze ein Übersetzungsprogramm, das mit einem MÜ-System verbunden ist. Danach post-editiere ich die Übersetzung.

c) Ich übersetze den Text und verwende ein MÜ-System, nur wenn ich es brauche (z. B. um einzelne Wörter / Sätze zu übersetzen).

d) Ich benutze gar kein MÜ-System. Bitte kurz begründen!

e) Ich habe eine andere Strategie für die Nutzung von MÜ-Systemen. Bitte kurz beschreiben!

2. **Hast Du Erfahrung mit der Kontrollierten Sprache (Controlled Language)?** *Mehrere Antworten sind möglich:*

a) Gar keine Erfahrung. Ich weiß nicht, was eine Kontrollierte Sprache ist.

b) Ich habe das Thema im Rahmen meines Studiums gelernt.

c) Ich habe nach bestimmten Regeln der Kontrollierten Sprache einen Übersetzungsauftrag (oder mehrere) beruflich oder als Übersetzungsübung im Laufe des Studiums angefertigt. Falls zutreffend, waren die Regeln vom Auftraggeber / Dozenten vorgegeben oder hast Du Dich dafür entschieden?

d) Ich beachte im Allgemeinen bei meiner technischen Übersetzung die Regeln der Kontrollierten Sprache. Falls zutreffend, warum?

e) Obwohl die Regeln der Kontrollierten Sprache mir bekannt sind, bevorzuge ich danach nicht zu übersetzen. Falls zutreffend, bitte kurz begründen!

f) Sonstiges: ……

Teil 2: Feedback zur Evaluation bzw. zur Testphase

1. Fandest Du die Evaluation lang / umfangreich?
2. Fandest Du die Evaluation interessant / langweilig? Falls langweilig, wann hat die Langweile eingesetzt, nach dem 10., 20., … Test?
3. Fandest Du den Aufbau der Evaluation verständlich / zu schwer?
4. Hast Du durch die Bewertung etwas Neues gelernt?
5. War in den DE- oder EN-Sätzen irgendetwas auffällig?
6. Kannst Du im Nachhinein erraten, worum es geht?

Anhang C: Datensatz

Quellen des analysierten Korpus

Ref.	Bezeichnung der Quelle	Hersteller
#1	Verlegeanweisung und Reinigungsverfahren für "EASY LIFT BAHNENWARE"	Halbmond Teppichwerke GmbH
#2	Gebrauchs- und Pflegeanleitung für keramikversiegeltes BERNDES-Kochgeschirr	Berndes Küche GmbH
#3	Betriebsanleitung des Feinstzerkleinerers "FZK-HPC-S3"	Hempe GmbH
#4	Handbuch der Konfigurationssoftware "SAUTER CASE VAV"	SAUTER Controls GmbH
#5	Gepäckregelung – Vermissen Sie Ihr Gepäck?	Deutsche Lufthansa AG
#6	Bedienungsanleitung des Milchschäumers "CREMIO"	Melitta Haushaltsprodukte GmbH & Co. KG
#7	Pflege und Bedienungsanleitung des Küchenmöbels von NOLTE	Nolte Küchen GmbH & Co. KG
#8	Technische Beschreibung und Bedienungsanleitung des Heimkino-Lautsprecher-Sets "System 10 THX Ultra 2"	Teufel GmbH
#9	Bedienungsanleitung der elektrischen Zitruspresse "ZP 40"	Rommelsbacher ElektroHausgeräte GmbH
#10	Betriebsanweisung der Haus- und Gartenpumpenserie "BP 3, BP 4, BP 5 und BP7"	Alfred Kärcher GmbH & Co. KG

Regel 1 – Für zitierte Oberflächentexte gerade Anführungszeichen verwenden

	KS	Sätze	Ref. Quelle
1	Vor	Das Tool bietet diese Korrekturen bei der Eingabe der Werte im Bereich **Übersicht** an.	#4
	Nach	Das Tool bietet diese Korrekturen bei der Eingabe der Werte im Bereich "**Übersicht**".	
2	Vor	Ist nur ein Gerät angeschlossen, so ist die Funktion **Punkt-zu-Punkt-Verbindung** zu wählen.	#4
	Nach	Ist nur ein Gerät angeschlossen, so ist die Funktion "**Punkt-zu-Punkt-Verbindung**" zu wählen.	
3	Vor	Durch Anklicken des Buttons **Zusatzinformation** werden die Angaben über das Netzwerk eingelesen.	#4
	Nach	Durch Anklicken des Buttons "**Zusatzinformation**" werden die Angaben über das Netzwerk eingelesen.	
4	Vor	Wenn eine korrekte Verbindung aufgebaut werden konnte, wird in der Statuszeile das Feld **Verbindung** grün.	#4
	Nach	Wenn eine korrekte Verbindung aufgebaut werden konnte, wird in der Statuszeile das Feld "**Verbindung**" grün.	
5	Vor	Im Reiter **Kommunikation BACnet** können die notwendigen Einstellungen vorgenommen werden.	#4
	Nach	Im Reiter "**Kommunikation BACnet**" können die notwendigen Einstellungen vorgenommen werden.	
6	Vor	Das Modul **ASV15** darf nur für seinen spezifizierten Einsatzzweck verwendet werden.	#4
	Nach	Das Modul "**ASV15**" darf nur für seinen spezifizierten Einsatzzweck verwendet werden.	
7	Vor	Im Abschnitt **Graphikeinstellungen** können Sie Einstellungen für die angezeigten Graphiken vornehmen.	#4
	Nach	Im Abschnitt "**Graphikeinstellungen**" können Sie Einstellungen für die angezeigten Graphiken vornehmen.	

8	Vor	Hierzu kann die Funktion **Upload vom Gerät** gewählt werden.	**#4**
	Nach	Hierzu kann die Funktion "**Upload vom Gerät**" gewählt werden.	
9	Vor	Überprüfen Sie die Adresse des Ports im **Geräte-Manager**.	**#4**
	Nach	Überprüfen Sie die Adresse des Ports im "**Geräte-Manager**".	
10	Vor	Wählen Sie die Option **Software von einer bestimmten Liste installieren**.	**#4**
	Nach	Wählen Sie die Option "**Software von einer bestimmten Liste installieren**".	
11	Vor	Klicken Sie auf der Startseite auf **Gerät konfigurieren**.	**#4**
	Nach	Klicken Sie auf der Startseite auf "**Gerät konfigurieren**".	
12	Vor	Klicken Sie auf **Netzwerk absuchen**, um die vorhandenen Geräte im Netzwerk anzuzeigen.	**#4**
	Nach	Klicken Sie auf "**Netzwerk absuchen**", um die vorhandenen Geräte im Netzwerk anzuzeigen.	
13	Vor	Unter dem Reiter **Einheiten** können Sie die verwendeten Einheiten einstellen.	**#4**
	Nach	Unter dem Reiter "**Einheiten**" können Sie die verwendeten Einheiten einstellen.	
14	Vor	Mit der Funktion **Geräteparameter ändern** können Sie die Parameter eines Gerätes ändern.	**#4**
	Nach	Mit der Funktion "**Geräteparameter ändern**" können Sie die Parameter eines Gerätes ändern.	
15	Vor	Die Funktion **Neue Adresse** ermöglicht die manuelle Vergabe einer neuen Netzwerkadresse.	**#4**
	Nach	Die Funktion "**Neue Adresse**" ermöglicht die manuelle Vergabe einer neuen Netzwerkadresse.	
16	Vor	Um ein neues Gerät zu konfigurieren, wählen Sie den Menüpunkt **Gerät konfigurieren** aus.	**#4**
	Nach	Um ein neues Gerät zu konfigurieren, wählen Sie den Menüpunkt "**Gerät konfigurieren**" aus.	

17	Vor	Die Firmware-Version wird in der Infozeile in der Spalte **Firmware-Version** angezeigt.	**#4**
	Nach	Die Firmware-Version wird in der Infozeile in der Spalte "**Firmware-Version**" angezeigt.	
18	Vor	Wenn Sie die Option **Zweites Gerät visualisieren** angewählt haben, können Sie die Parameter bestimmen.	**#4**
	Nach	Wenn Sie die Option "**Zweites Gerät visualisieren**" angewählt haben, können Sie die Parameter bestimmen.	
19	Vor	Im Eingabefenster **Projektdaten** werden Informationen zur Lokalisierung des Reglers eingegeben.	**#4**
	Nach	Im Eingabefenster "**Projektdaten**" werden Informationen zur Lokalisierung des Reglers eingegeben.	
20	Vor	Auf der Seite **Einstellungen** sind alle notwendigen Parameter zur Optimierung des Regelkreises zusammengefasst.	**#4**
	Nach	Auf der Seite "**Einstellungen**" sind alle notwendigen Parameter zur Optimierung des Regelkreises zusammengefasst.	
21	Vor	Die zwei aktivierbaren Raumdruck-Sollwerte sind im Menü **Raumdruck** definiert.	**#4**
	Nach	Die zwei aktivierbaren Raumdruck-Sollwerte sind im Menü "**Raumdruck**" definiert.	
22	Vor	Wählen Sie danach die Option **Software automatisch installieren**.	**#4**
	Nach	Wählen Sie danach die Option "**Software automatisch installieren**".	
23	Vor	Wählen Sie die Option **Auswahl nach Anwendungs-Code** aus.	**#4**
	Nach	Wählen Sie die Option "**Auswahl nach Anwendungs-Code**" aus.	
24	Vor	Mit der Funktion **Anwendung neu konfigurieren**, kann dem Gerät eine neue Anwendung zugewiesen werden.	**#4**

	Nach	Mit der Funktion "**Anwendung neu konfigurieren**", kann dem Gerät eine neue Anwendung zugewiesen werden.

Regel 2 – Funktionsverbgefüge vermeiden

	KS	Sätze	Ref. Quelle
1	Vor	Der Bediener darf erst die Maschine **in Betrieb nehmen**, wenn er die Betriebsanleitung gelesen hat.	**#3**
	Nach	Der Bediener darf erst die Maschine **starten**, wenn er die Betriebsanleitung gelesen hat.	
2	Vor	Der Hersteller **übernimmt** keine **Haftung** für Schäden, die durch nicht bestimmungsgemäßen Gebrauch entstanden sind.	**#8**
	Nach	Der Hersteller **haftet** nicht für Schäden, die durch nicht bestimmungsgemäßen Gebrauch entstanden sind.	
3	Vor	**Steht** die Maschine nicht **im Einsatz**, den Hauptschalter auf "0" setzen.	**#3**
	Nach	**Wird** die Maschine nicht **verwendet**, den Hauptschalter auf "0" setzen.	
4	Vor	Wird diese Regel nicht beachtet, kann der Motor **Schaden nehmen**.	**#9**
	Nach	Wird diese Regel nicht beachtet, kann der Motor **beschädigt werden**.	
5	Vor	Die Fleck**behandlung** muss so schnell wie möglich **durchgeführt** werden.	**#1**
	Nach	Die Flecken müssen so schnell wie möglich **behandelt** werden.	
6	Vor	Der Kompaktregler **setzt sich mit** der neuen Netzwerkadresse **in Verbindung**.	**#4**
	Nach	Der Kompaktregler **verbindet sich mit** der neuen Netzwerkadresse.	

7	Vor	Der Navigationsbaum **stellt** alle vorhandenen Seiten der Konfigurierung **zur Verfügung**.	**#4**
	Nach	Der Navigationsbaum **stellt** alle vorhandenen Seiten der Konfigurierung **bereit**.	
8	Vor	Die vorderen Schutzgitter **bieten** den empfindlichen Lautsprechermembranen **Schutz**.	**#8**
	Nach	Die vorderen Schutzgitter **schützen** die empfindlichen Lautsprechermembranen.	
9	Vor	Küchenmöbel**montagen** dürfen nur von geschulten Fachleuten **durchgeführt** werden.	**#7**
	Nach	Küchenmöbel dürfen nur von geschulten Fachleuten **montiert** werden.	
10	Vor	Die **Reinigung** der Küchenmöbel sollten Sie mit einem leicht feuchten Tuch **vornehmen**.	**#7**
	Nach	Sie sollten die Küchenmöbel mit einem leicht feuchten Tuch **reinigen**.	
11	Vor	Systemlösungen, die in öffentlichen Bereichen **zum Einsatz kommen**, müssen auf ihre Brennklasse geprüft werden.	**#1**
	Nach	Systemlösungen, die in öffentlichen Bereichen **eingesetzt werden**, müssen auf ihre Brennklasse geprüft werden.	
12	Vor	Es **liegt in der Verantwortung** des Planers, aufeinander abgestimmte Produkte einzusetzen.	**#1**
	Nach	Der Planer **ist dafür verantwortlich**, aufeinander abgestimmte Produkte einzusetzen.	
13	Vor	Somit kann die Fluggesellschaft nicht garantieren, dass die Gepäckregeln immer **zur Anwendung kommen**.	**#5**
	Nach	Somit kann die Fluggesellschaft nicht garantieren, dass die Gepäckregeln immer **angewendet werden**.	
14	Vor	Die **Abwicklung** von Garantieleistungen **erfolgt** über die lokale Service-Hotline.	**#6**
	Nach	Die Garantieleistungen **werden** über die lokale Service-Hotline **abgewickelt**.	

15	Vor	Das Gerät darf nicht **in Betrieb genommen werden**, wenn es sichtbare Schäden aufweist.	**#9**
	Nach	Das Gerät darf nicht **eingeschaltet werden**, wenn es sichtbare Schäden aufweist.	
16	Vor	Die Höhen**verstellung** der Fronten können Sie mittels eines Schraubendrehers **vornehmen**.	**#7**
	Nach	Die Höhe der Fronten können Sie mittels eines Schraubendrehers **verstellen**.	
17	Vor	Im oberen Abschnitt können Sie **Einstellungen** für die angezeigten Module **vornehmen**.	**#4**
	Nach	Im oberen Abschnitt können Sie die angezeigten Module **einstellen**.	
18	Vor	Ist der Wert abweichend, kann eine **Korrektur vorgenommen werden**.	**#4**
	Nach	Ist der Wert abweichend, kann er **korrigiert werden**.	
19	Vor	Auf der Startseite **stehen** die folgenden Funktionen zur Auswahl **zur Verfügung**.	**#4**
	Nach	Auf der Startseite **sind** die folgenden Funktionen zur Auswahl **vorhanden**.	
20	Vor	Das Schutzmodul der Maschine darf nicht **außer Betrieb gesetzt werden**.	**#3**
	Nach	Das Schutzmodul der Maschine darf nicht **abschaltet werden**.	
21	Vor	Die Maschine **ist** täglich 8 Stunden **in Betrieb**.	**#10**
	Nach	Die Maschine **arbeitet** täglich 8 Stunden.	
22	Vor	Sie haben die Möglichkeit, einen **Antrag** auf Verlängerung der Frist zu **stellen**.	**#10**
	Nach	Sie haben die Möglichkeit, eine Verlängerung der Frist zu **beantragen**.	
23	Vor	Nach Ihrer Registrierung im Programm können Sie aus den Leistungen eine **Auswahl treffen**.	**#10**
	Nach	Nach Ihrer Registrierung im Programm können Sie aus den Leistungen **wählen**.	

24	Vor	Die Kunden **zeigen** vor allem **Interesse an** den technischen Neuentwicklungen.	**#10**
	Nach	Die Kunden **interessieren sich** vor allem **für** die technischen Neuentwicklungen.	

Regel 3 – Konditionalsätze mit ‚Wenn' einleiten

	KS	Sätze	Ref. Quelle
1	Vor	**Ist** die Seriennummer des Gerätes bekannt, kann im Feld Seriennummer diese Nummer eingegeben werden.	**#4**
	Nach	**Wenn** die Seriennummer des Gerätes bekannt **ist**, kann im Feld Seriennummer diese Nummer eingegeben werden.	
2	Vor	**Steht** die Maschine nicht im Einsatz, den Hauptschalter auf "0" setzen.	**#3**
	Nach	**Wenn** die Maschine nicht im Einsatz **steht**, den Hauptschalter auf "0" setzen.	
3	Vor	**Schließt** der Kontaktschalter, so wird der Raumdruck-Sollwert aktiv.	**#4**
	Nach	**Wenn** der Kontaktschalter **schließt**, wird der Raumdruck-Sollwert aktiv.	
4	Vor	**Werden** die vordefinierten Werte **verändert**, so erfolgt die Umrechnung automatisch.	**#4**
	Nach	**Wenn** die vordefinierten Werte **verändert werden**, erfolgt die Umrechnung automatisch.	
5	Vor	**Ist** diese Zeit **erreicht**, muss das Gerät für 2 Minuten abkühlen.	**#9**
	Nach	**Wenn** diese Zeit **erreicht ist**, muss das Gerät für 2 Minuten abkühlen.	
6	Vor	**Wird** der zweite Anschluss als Eingang **konfiguriert**, so muss der Sollwert angepasst werden.	**#4**

	Nach	**Wenn** der zweite Anschluss als Eingang **konfiguriert wird**, muss der Sollwert angepasst werden.	
7	Vor	**Ist** die importierte Firmware-Version älter als die installierte Version, erhält der Benutzer eine Warnmeldung.	**#4**
	Nach	**Wenn** die importierte Firmware-Version älter als die installierte Version **ist**, erhält der Benutzer eine Warnmeldung.	
8	Vor	**Ist** nur ein Gerät **angeschlossen**, so ist die Funktion PP zu wählen.	**#4**
	Nach	**Wenn** nur ein Gerät **angeschlossen ist**, ist die Funktion PP zu wählen.	
9	Vor	**Wählt** man einen bestimmten Zeichensatz als Standardwert, so wird dieser Zeichensatz in allen Stationen verwendet.	**#4**
	Nach	**Wenn** man einen bestimmten Zeichensatz als Standardwert **wählt**, wird dieser Zeichensatz in allen Stationen verwendet.	
10	Vor	**Steht** ein normierter Faktor zur Verfügung, kann dieser Faktor direkt in der Eingabemaske eingegeben werden.	**#4**
	Nach	**Wenn** ein normierter Faktor zur Verfügung **steht**, kann dieser Faktor direkt in der Eingabemaske eingegeben werden.	
11	Vor	**Wird** diese Regel nicht **beachtet**, kann der Motor Schaden nehmen.	**#9**
	Nach	**Wenn** diese Regel nicht **beachtet wird**, kann der Motor Schaden nehmen.	
12	Vor	**Ist** ein mehrstufiges Modul **parametriert**, so sind die externen Kontakte zu verriegeln.	**#4**
	Nach	**Wenn** ein mehrstufiges Modul **parametriert ist**, sind die externen Kontakte zu verriegeln.	
13	Vor	**Sind** mehrere Geräte im Netzwerk vorhanden, so ist die Adresse des gewünschten Gerätes auszuwählen.	**#4**
	Nach	**Wenn** mehrere Geräte im Netzwerk vorhanden **sind**, ist die Adresse des gewünschten Gerätes auszuwählen.	

14	Vor	**Erfolgt** die Zahlung nicht, kann der Anbieter Ersatz eines eventuell entstandenen Schadens verlangen.	**#2**
	Nach	**Wenn** die Zahlung nicht erfolgt, kann der Anbieter Ersatz eines eventuell entstandenen Schadens verlangen.	
15	Vor	**Ist** der c-Faktor mit einer anderen Luftdichte **angegeben worden**, so ist diese Luftdichte einzutragen.	**#4**
	Nach	**Wenn** der c-Faktor mit einer anderen Luftdichte **angegeben worden ist**, ist diese Luftdichte einzutragen.	
16	Vor	**Ist** das Gerät oder das Netzkabel **beschädigt**, sofort den Netzstecker herausziehen.	**#9**
	Nach	**Wenn** das Gerät oder das Netzkabel **beschädigt ist**, sofort den Netzstecker herausziehen.	
17	Vor	**Werden** Geräte in einem anderen Land **gekauft**, werden Garantieleistungen nur in diesem Land erbracht.	**#6**
	Nach	**Wenn** Geräte in einem anderen Land **gekauft werden**, werden Garantieleistungen nur in diesem Land erbracht.	
18	Vor	**Tritt** eine Preisänderung **ein**, so gilt der neue Preis am Tag der Lieferung.	**#2**
	Nach	**Wenn** eine Preisänderung **eintritt**, gilt der neue Preis am Tag der Lieferung.	
19	Vor	**Veräußert** der Besteller die gelieferte, unbezahlte Ware, so tritt er dem Lieferer alle Ansprüche ab.	**#2**
	Nach	**Wenn** der Besteller die gelieferte, unbezahlte Ware **veräußert**, tritt er dem Lieferer alle Ansprüche ab.	
20	Vor	**Wurde** die Maschine mit verdeckten Beschädigungen **angeliefert**, so verständigen Sie unverzüglich den Lieferanten.	**#3**
	Nach	**Wenn** die Maschine mit verdeckten Beschädigungen **angeliefert wurde**, verständigen Sie unverzüglich den Lieferanten.	
21	Vor	**Liegt** die Regelabweichung innerhalb der x-Zone, so bleibt das erste Modul stehen.	**#4**
	Nach	**Wenn** die Regelabweichung innerhalb der x-Zone **liegt**, bleibt das erste Modul stehen.	

22	Vor	**Wird** der Startbildschirm nicht **angezeigt**, war die Installation wahrscheinlich fehlerhaft.	**#4**
	Nach	**Wenn** der Startbildschirm nicht **angezeigt wird**, war die Installation wahrscheinlich fehlerhaft.	
23	Vor	**Erscheint** eine Fehlermeldung, war die Konfigurierung wahrscheinlich unvollständig.	**#4**
	Nach	**Wenn** eine Fehlermeldung **erscheint**, war die Konfigurierung wahrscheinlich unvollständig.	
24	Vor	**Werden** die Standardwerte der Maschine **überschritten**, führt die Überlastung zum automatischen Abschalten der Maschine.	**#3**
	Nach	**Wenn** die Standardwerte der Maschine **überschritten werden**, führt die Überlastung zum automatischen Abschalten der Maschine.	

Regel 4 – Eindeutige pronominale Bezüge verwenden

	KS	Sätze	Ref. Quelle
1	Vor	Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, **ihn** rückstandslos zu entfernen.	**#1**
	Nach	Je früher ein Fleck behandelt wird, umso größer ist die Wahrscheinlichkeit, **den Fleck** rückstandslos zu entfernen.	
2	Vor	Sofern auf der Oberfläche alte Klebereste anhaften, sind **diese** vollständig zu entfernen.	**#1**
	Nach	Sofern auf der Oberfläche alte Klebereste anhaften, sind **diese Kleberreste** vollständig zu entfernen.	
3	Vor	Um die Startlinie festzustellen, kann mit Kreide **diese** markiert werden.	**#1**
	Nach	Um die Startlinie festzustellen, kann mit Kreide **diese Linie** markiert werden.	

4	Vor	Bei einer Überhitzung bestehen keine gesundheitlichen Risiken; **diese** werden eher von verbranntem Öl oder Fett verursacht.	**#2**
	Nach	Bei einer Überhitzung bestehen keine gesundheitlichen Risiken; **gesundheitliche Risiken** werden eher von verbranntem Öl oder Fett verursacht.	
5	Vor	Fettreste müssen vollständig abgewaschen werden, da sich **diese** ansonsten in der Pfanne einbrennen können.	**#2**
	Nach	Fettreste müssen vollständig abgewaschen werden, da sich **diese Reste** ansonsten in der Pfanne einbrennen können.	
6	Vor	Halten Sie die Schaltschränke stets verschlossen, wenn **diese** unbeaufsichtigt sind.	**#3**
	Nach	Halten Sie die Schaltschränke stets verschlossen, wenn **die Schaltschränke** unbeaufsichtigt sind.	
7	Vor	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **diese** warten.	**#3**
	Nach	Nur Elektrofachkräfte dürfen Zugang zur Elektrik der Maschine haben und **die Maschine** warten.	
8	Vor	Wenn Störungen an der elektrischen Energieversorgung der Maschine auftreten, ist **diese** sofort mit dem Hauptschalter auszuschalten!	**#3**
	Nach	Wenn Störungen an der elektrischen Energieversorgung der Maschine auftreten, ist **die Maschine** sofort mit dem Hauptschalter auszuschalten!	
9	Vor	Tab 7 und Tab 8 zeigen die verfügbaren Anwendungen und **deren** unterschiedlichen Default-Parameter.	**#4**
	Nach	Tab 7 und Tab 8 zeigen die verfügbaren Anwendungen und **die** unterschiedlichen Default-Parameter **dieser Anwendungen**.	
10	Vor	Durch Anklicken des Buttons "Zusatzinformation" wird **sie** über das Netzwerk eingelesen.	**#4**
	Nach	Durch Anklicken des Buttons "Zusatzinformation" wird **die Zusatzinformation** über das Netzwerk eingelesen.	

11	Vor	Wählt man einen bestimmten Zeichensatz als Standardwert, so wird **der** in allen Stationen verwendet.	**#4**
	Nach	Wählt man einen bestimmten Zeichensatz als Standardwert, so wird **dieser Zeichensatz** in allen Stationen verwendet.	
12	Vor	Steht ein normierter Faktor zur Verfügung, kann direkt mit **diesem** in der Eingabemaske gearbeitet werden.	**#4**
	Nach	Steht ein normierter Faktor zur Verfügung, kann direkt mit **diesem Faktor** in der Eingabemaske gearbeitet werden.	
13	Vor	Veräußert der Besteller die gelieferte, unbezahlte Ware, so tritt **er** dem Lieferer alle Ansprüche ab.	**#2**
	Nach	Veräußert der Besteller die gelieferte, unbezahlte Ware, so tritt **der Besteller** dem Lieferer alle Ansprüche ab.	
14	Vor	Der Bediener darf erst die Maschine in Betrieb nehmen, wenn **er** die Betriebsanleitung gelesen hat.	**#3**
	Nach	Der Bediener darf erst die Maschine in Betrieb nehmen, wenn **der Bediener** die Betriebsanleitung gelesen hat.	
15	Vor	Wenn Sie einen Schaden erst zu Hause feststellen, melden Sie **ihn** innerhalb von sieben Tagen schriftlich.	**#5**
	Nach	Wenn Sie einen Schaden erst zu Hause feststellen, melden Sie **diesen Schaden** innerhalb von sieben Tagen schriftlich.	
16	Vor	Um den hohen Wert der Maschine über Jahre zu erhalten, sollten Sie **sie** richtig pflegen.	**#7**
	Nach	Um den hohen Wert der Maschine über Jahre zu erhalten, sollten Sie **die Maschine** richtig pflegen.	
17	Vor	Ziehen Sie sofort den Netzstecker und betreiben Sie das Gerät nicht, wenn **dessen** Gehäuse defekt ist.	**#8**
	Nach	Ziehen Sie sofort den Netzstecker und betreiben Sie das Gerät nicht, wenn **das Gehäuse** defekt ist.	
18	Vor	Heben Sie die Bedienungsanleitung gut auf und übergeben Sie **sie** auch an einen möglichen Nachbesitzer.	**#8**

	Nach	Heben Sie die Bedienungsanleitung gut auf und übergeben Sie **die Bedienungsanleitung** auch an einen möglichen Nachbesitzer.	
19	Vor	Jeder Verbraucher ist gesetzlich verpflichtet, alte Geräte bei einer Sammelstelle abzugeben, damit **sie** recycelt werden können.	**#8**
	Nach	Jeder Verbraucher ist gesetzlich verpflichtet, alte Geräte bei einer Sammelstelle abzugeben, damit **diese Geräte** recycelt werden können.	
20	Vor	Drehen Sie den Saftbehälter im Uhrzeigersinn bis **er** hörbar einrastet.	**#9**
	Nach	Drehen Sie den Saftbehälter im Uhrzeigersinn bis **der Saftbehälter** hörbar einrastet.	
21	Vor	Ist der c-Faktor mit einer anderen Luftdichte angegeben worden, so ist **dieser** im Feld "Luftdichte" einzutragen.	**#4**
	Nach	Ist der c-Faktor mit einer anderen Luftdichte angegeben worden, so ist **dieser Faktor** im Feld "Luftdichte" einzutragen.	
22	Vor	Durch Anklicken der entsprechenden Seite wird **diese** aktiv.	**#4**
	Nach	Durch Anklicken der entsprechenden Seite wird **die Seite** aktiv.	
23	Vor	Durch Klick auf die angezeigte Adresse kann **diese** im Menü konfiguriert werden.	**#4**
	Nach	Durch Klick auf die angezeigte Adresse kann **die Adresse** im Menü konfiguriert werden.	
24	Vor	Überprüfen Sie die Zuweisung des Ports im Geräte-Manager und stellen Sie **diesen** ggf. um.	**#4**
	Nach	Überprüfen Sie die Zuweisung des Ports im Geräte-Manager und stellen Sie **diesen Port** ggf. um.	

Regel 5 – Partizipial-konstruktionen vermeiden

	KS	Sätze	Ref. Quelle
1	Vor	Das Gerät nur an **eine vorschriftsmäßig installierte Steckdose** anschließen.	**#9**
	Nach	Das Gerät nur an **eine Steckdose** anschließen, **die vorschriftsmäßig installiert ist**.	
2	Vor	Kunststoffverpackungen in **die dafür vorgesehenen Entsorgungsbehälter** geben.	**#9**
	Nach	Kunststoffverpackungen in **die Entsorgungsbehälter** geben, **die dafür vorgesehen sind**.	
3	Vor	**Speziell auf diese Lautsprecher abgestimmtes Zubehör** erhalten Sie in unserem Webshop.	**#8**
	Nach	**Zubehör, das speziell auf diese Lautsprecher abgestimmt ist,** erhalten Sie in unserem Webshop.	
4	Vor	Beachten Sie **die zusätzlich ausgehändigten Bedienungsanleitungen** der Einbaukomponenten.	**#7**
	Nach	Beachten Sie **die Bedienungsanleitungen** der Einbaukomponenten, **die zusätzlich ausgehändigt wurden**.	
5	Vor	**Alle von uns eingesetzten Schubkästen und Einlegeböden** sind hochwertige Produkte.	**#7**
	Nach	**Alle Schubkästen und Einlegeböden, die von uns eingesetzt werden,** sind hochwertige Produkte.	
6	Vor	Mittels **des von Ihnen ausgefüllten Formulars** wird eine Suche durchgeführt.	**#5**
	Nach	Mittels **des Formulars, das von Ihnen ausgefüllt wird**, wird eine Suche durchgeführt.	
7	Vor	**Die für Ihren Flug erlaubte Freigepäckmenge** ist auf Ihrem Flugschein angegeben.	**#5**
	Nach	**Die Freigepäckmenge, die für Ihren Flug erlaubt ist,** ist auf Ihrem Flugschein angegeben.	
8	Vor	Hierzu kann die Funktion "Upload" gewählt werden, wodurch **die im Gerät gespeicherten Daten** geladen werden.	**#4**

	Nach	Hierzu kann die Funktion "Upload" gewählt werden, wodurch **die Daten, die im Gerät gespeichert sind**, geladen werden.	
9	Vor	Die eingegebene Netzwerkadresse bezieht sich immer auf **den in der Infozeile angezeigten Gerätetyp**.	**#4**
	Nach	Die eingegebene Netzwerkadresse bezieht sich immer auf **den Gerätetyp, der in der Infozeile angezeigt ist**.	
10	Vor	Das Gerät verbindet sich mit **der neu gewählten Netzwerkadresse**.	**#4**
	Nach	Das Gerät verbindet sich mit **der Netzwerkadresse, die neu gewählt wird**.	
11	Vor	Durch Eingabe **der mit einem roten Sternchen gekennzeichneten Parameter** erfolgt die minimale Konfigurierung.	**#4**
	Nach	Durch Eingabe **der Parameter, die mit einem roten Sternchen gekennzeichnet sind**, erfolgt die minimale Konfigurierung.	
12	Vor	In den Einstellungen können Sie **die grafisch angezeigten Werte** in eine CSV-Datei speichern.	**#4**
	Nach	In den Einstellungen können Sie **die Werte, die grafisch angezeigt werden**, in eine CSV-Datei speichern.	
13	Vor	Dank **dem im System integrierten zweiten C-Modul** werden weitere Anwendungen unterstützt.	**#4**
	Nach	Dank **dem zweiten C-Modul, das im System integriert ist**, werden weitere Anwendungen unterstützt.	
14	Vor	Der Digitaleingang dient zur Steuerung **des am Analogeingang angelegten Sollwertes**.	**#4**
	Nach	Der Digitaleingang dient zur Steuerung **des Sollwertes, der am Analogeingang angelegt ist**.	
15	Vor	**Die in der Betriebsanleitung angegebenen Fristen** für wiederkehrende Prüfungen sind einzuhalten.	**#3**
	Nach	**Die Fristen** für wiederkehrende Prüfungen, **die in der Betriebsanleitung angegeben sind**, sind einzuhalten.	

16	Vor	Die Daten der Menüs werden für **die neu ausgewählte Anwendung** beibehalten.	**#4**
	Nach	Die Daten der Menüs werden für **die Anwendung** beibehalten, **die neu ausgewählt worden ist**.	
17	Vor	**Alle darüberhinausgehenden Ansprüche** sind ausdrücklich von der Garantie ausgenommen.	**#2**
	Nach	**Alle Ansprüche**, **die darüber hinausgehen**, sind ausdrücklich von der Garantie ausgenommen.	
18	Vor	Bei **fristgerecht erfolgten berechtigten Mängelrügen** ist der Lieferer zu einer kostenlosen Ersatzlieferung verpflichtet.	**#2**
	Nach	Bei **berechtigten Mängelrügen**, **die fristgerecht erfolgen**, ist der Lieferer zu einer kostenlosen Ersatzlieferung verpflichtet.	
19	Vor	Erfolgt die Zahlung nicht, kann der Anbieter Ersatz **eines eventuell entstandenen Schadens** verlangen.	**#2**
	Nach	Erfolgt die Zahlung nicht, kann der Anbieter Ersatz **eines Schadens** verlangen, **der eventuell entsteht**.	
20	Vor	Die produktspezifischen Betriebsanleitungen sind auf **der zusätzlich gelieferten CD** dokumentiert.	**#3**
	Nach	Die produktspezifischen Betriebsanleitungen sind auf **der CD** dokumentiert, **die zusätzlich geliefert wurde**.	
21	Vor	**Die in der Betriebsanleitung enthaltenen Sicherheitshinweise** sind mit dem allgemeinen Gefahrensymbol gekennzeichnet.	**#3**
	Nach	**Die Sicherheitshinweise, die in der Betriebsanleitung enthalten sind**, sind mit dem allgemeinen Gefahrensymbol gekennzeichnet.	
22	Vor	**Die in den Bedienungsanweisungen der eingebauten Geräte vorgeschriebenen Gebrauchsbedingungen** müssen strikt eingehalten werden.	**#3**
	Nach	**Die Gebrauchsbedingungen**, **die in den Bedienungsanweisungen der eingebauten Geräte vorgeschrieben sind**, müssen strikt eingehalten werden.	

23	Vor	**Für hieraus resultierende Schäden** haftet allein der Betreiber der Maschine.	#3
	Nach	**Für Schäden, die hieraus resultieren**, haftet allein der Betreiber der Maschine.	
24	Vor	**Die für die Maschine benötigten Werkzeuge** sind im Lieferumfang nicht enthalten.	#3
	Nach	**Die Werkzeuge, die für die Maschine benötigt werden**, sind im Lieferumfang nicht enthalten.	

Regel 6 – Passiv vermeiden

	KS	Sätze	Ref. Quelle
1	Vor	Bei der Arbeit mit elektrischen Geräten **sollte** stets ein Sicherheitsstecker **verwendet werden**.	#3
	Nach	Bei der Arbeit mit elektrischen Geräten **verwenden Sie** stets einen Sicherheitsstecker.	
2	Vor	Ist die Seriennummer des Gerätes bekannt, **kann** im Feld Seriennummer diese Nummer **eingegeben werden**.	#4
	Nach	Ist die Seriennummer des Gerätes bekannt, **können Sie** im Feld Seriennummer diese Nummer **eingeben**.	
3	Vor	Wenn eine korrekte Verbindung **aufgebaut werden konnte**, wird in der Statuszeile das Feld Verbindung grün.	#4
	Nach	Wenn **Sie** eine korrekte Verbindung **aufbauen konnten**, wird in der Statuszeile das Feld Verbindung grün.	
4	Vor	Im Reiter Kommunikation **können** die notwendigen Einstellungen zur Kommunikation über das IP-Netzwerk **vorgenommen werden**.	#4
	Nach	Im Reiter Kommunikation **können Sie** die notwendigen Einstellungen zur Kommunikation über das IP-Netzwerk **vornehmen**.	

5	Vor	Reparaturen **dürfen** nur von autorisierten Fachbetrieben **durchgeführt werden**.	**#9**
	Nach	Nur autorisierte Fachbetriebe **dürfen** die Reparaturen **durchführen**.	
6	Vor	Das Gerät stoppt, sobald der Druck auf den Presskegel **vermindert wird**.	**#9**
	Nach	Das Gerät stoppt, sobald **Sie** den Druck auf den Presskegel **vermindern**.	
7	Vor	Flecken **sollten** so schnell wie möglich **behandelt werden**.	**#1**
	Nach	Flecken **sollten Sie** so schnell wie möglich **behandeln**.	
8	Vor	Nach Ablauf des Spülprogramms **sollte** der Geschirrspüler nicht sofort **geöffnet werden**.	**#7**
	Nach	Nach Ablauf des Spülprogramms **sollten Sie** den Geschirrspüler nicht sofort **öffnen**.	
9	Vor	Die akustischen Signale **können** je nach Gerät **umprogrammiert werden**.	**#7**
	Nach	Die akustischen Signale **können Sie** je nach Gerät **umprogrammieren**.	
10	Vor	Achten Sie darauf, dass das Infrarotlicht nicht durch Gegenstände **behindert wird**.	**#8**
	Nach	Achten Sie darauf, dass keine Gegenstände das Infrarotlicht **behindern**.	
11	Vor	Das Modul XY **darf** nur für seinen spezifizierten Einsatzzweck **verwendet werden**.	**#4**
	Nach	**Sie dürfen** das Modul XY nur für seinen spezifizierten Einsatzzweck **verwenden**.	
12	Vor	Durch diese Öffnung **kann** der Stecker mit dem Regler **verbunden werden**.	**#4**
	Nach	Durch diese Öffnung **können Sie** den Stecker mit dem Regler **verbinden**.	
13	Vor	Um mit einem anderen Gerätetyp zu kommunizieren, **muss** zuerst die Gerätenummer **eingegeben werden**.	**#4**

	Nach	Um mit einem anderen Gerätetyp zu kommunizieren, **müssen Sie** zuerst die Gerätenummer **eingeben**.	
14	Vor	In dem Drop-down-Menü **kann** die gewünschte IP-Adresse **gewählt werden**.	**#4**
	Nach	In dem Drop-down-Menü **können Sie** die gewünschte IP-Adresse **wählen**.	
15	Vor	Sinkt der Druck unter ca. 1,3 bar, **wird** die Pumpe **gestartet**.	**#10**
	Nach	Sinkt der Druck unter ca. 1,3 bar, **startet** die Pumpe.	
16	Vor	Bei der Verwendung von nur einem Ausgang **kann** der zweite Ausgang **verschlossen werden**.	**#10**
	Nach	Bei der Verwendung von nur einem Ausgang **können Sie** den zweiten Ausgang **verschließen**.	
17	Vor	Der EIN/AUS-Schalter **kann** komfortabel mit dem Fuß **betätigt werden**.	**#10**
	Nach	**Sie können** den EIN/AUS-Schalter komfortabel mit dem Fuß **betätigen**.	
18	Vor	Das Programm **wird** vom Hersteller wie folgt **eingestellt**.	**#3**
	Nach	Der Hersteller **stellt** das Programm wie folgt **ein**.	
19	Vor	Schutzvorrichtungen **dürfen** nur nach Stillstand der Maschine **entfernt werden**.	**#3**
	Nach	**Sie dürfen** die Schutzvorrichtungen nur nach Stillstand der Maschine **entfernen**.	
20	Vor	Die Maschine **darf** nur mithilfe eines Gabelstaplers **angehoben werden**.	**#3**
	Nach	**Sie dürfen** die Maschine nur mithilfe eines Gabelstaplers **anheben**.	
21	Vor	Die Transportösen **müssen** nach dem Transport der Maschine **demontiert werden**.	**#3**
	Nach	**Sie müssen** die Transportösen nach dem Transport der Maschine **demontieren**.	
22	Vor	Die Laufrollen **müssen** nach dem Aufstellen der Maschine **verriegelt werden**.	**#3**

	Nach	**Sie müssen** die Laufrollen nach dem Aufstellen der Maschine **verriegeln**.	
23	Vor	Die Konfigurierung des Moduls **kann** in eine Datei **exportiert werden**.	**#4**
	Nach	**Sie können** die Konfigurierung des Moduls in eine Datei **exportieren**.	
24	Vor	Alle Konfigurationsdaten **können** mithilfe der Druckfunktion in eine Datei **gedruckt werden**.	**#4**
	Nach	**Sie können** alle Konfigurationsdaten mithilfe der Druckfunktion in eine Datei **drucken**.	

Regel 7 – Konstruktionen mit ‚sein + zu + Infinitiv' vermeiden

	KS	Sätze	Ref. Quelle
1	Vor	Mängel **sind** unverzüglich **zu melden**.	**#1**
	Nach	**Melden Sie** etwaige Mängel unverzüglich.	
2	Vor	Die Herstelleranweisungen **sind** stets **zu beachten**.	**#1**
	Nach	**Beachten Sie** stets die Herstelleranweisungen.	
3	Vor	Festgestellte Schäden **sind** sofort **zu beheben**.	**#3**
	Nach	**Beheben Sie** festgestellte Schäden sofort.	
4	Vor	Das Kaufdatum **ist** durch eine Kaufquittung **zu belegen**.	**#6**
	Nach-KS	**Belegen Sie** das Kaufdatum durch eine Kaufquittung.	
5	Vor	Die Bedienungsanleitung **ist** im Falle des Verlustes **zu ersetzen**.	**#7**
	Nach	**Ersetzen Sie** die Bedienungsanleitung im Falle des Verlustes.	
6	Vor	Ist nur ein Gerät angeschlossen, so **ist** die Funktion PP **zu wählen**.	**#4**

	Nach	Ist nur ein Gerät angeschlossen, so **wählen Sie** die Funktion PP.	
7	Vor	Die Teppichböden **sind** entsprechend den Liefer- und Zahlungsbedingungen **zu prüfen**.	**#1**
	Nach	**Prüfen Sie** die Teppichböden entsprechend den Liefer- und Zahlungsbedingungen.	
8	Vor	Bei Funktionsstörungen **ist** die Maschine sofort **auszuschalten**.	**#3**
	Nach	Bei Funktionsstörungen **schalten Sie** die Maschine sofort **aus**.	
9	Vor	Das Bedienungspersonal **ist** über das Problem vor dem Reparaturbeginn **zu informieren**.	**#3**
	Nach	**Informieren Sie** das Bedienungspersonal über das Problem vor dem Reparaturbeginn.	
10	Vor	Die Maschine **ist** gegen das Einschalten durch Unbefugte **zu sichern**.	**#3**
	Nach	**Sichern Sie** die Maschine gegen das Einschalten durch Unbefugte.	
11	Vor	Folgende Reinigungsarbeiten der Schneidwerkzeuge **sind** täglich nach Produktionsende **durchzuführen**.	**#3**
	Nach	**Führen Sie** folgende Reinigungsarbeiten der Schneidwerkzeuge täglich nach Produktionsende **durch**.	
12	Vor	Sofern auf der Oberfläche alte Kleberreste anhaften, **sind** diese vollständig **zu entfernen**.	**#1**
	Nach	Sofern auf der Oberfläche alte Kleberreste anhaften, **entfernen Sie** diese vollständig.	
13	Vor	Die vorgeschriebenen Fristen für wiederkehrende Inspektionen **sind einzuhalten**.	**#3**
	Nach	**Halten Sie** die vorgeschriebenen Fristen für wiederkehrende Inspektionen **ein**.	
14	Vor	Sämtliche Wartungsarbeiten **sind** nach Betriebsanleitung des Herstellers **auszuführen**.	**#3**
	Nach	**Führen Sie** sämtliche Wartungsarbeiten nach Betriebsanleitung des Herstellers **aus**.	

15	Vor	Hierzu **ist** die Funktion Manueller Betrieb im Bereich Servicefunktionen **zu wählen**.	**#4**
	Nach	Hierzu **wählen Sie** die Funktion Manueller Betrieb im Bereich Servicefunktionen.	
16	Vor	Ist ein mehrstufiges Modul parametriert, so **sind** die externen Kontakte **zu verriegeln**.	**#4**
	Nach	Ist ein mehrstufiges Modul parametriert, **verriegeln Sie** die externen Kontakte.	
17	Vor	Um die Verbindung mit dem Regler herzustellen, **ist** die Verschlusskappe **zu öffnen**.	**#4**
	Nach	Um die Verbindung mit dem Regler herzustellen, **öffnen Sie** die Verschlusskappe.	
18	Vor	Vor der Parametrierung **ist** der Regler **zu konfigurieren**.	**#4**
	Nach	Vor der Parametrierung **konfigurieren Sie** den Regler.	
19	Vor	Ist der c-Faktor mit einer anderen Luftdichte angegeben worden, so **ist** diese Luftdichte **einzutragen**.	**#4**
	Nach	Ist der c-Faktor mit einer anderen Luftdichte angegeben worden, **tragen Sie** diese Luftdichte **ein**.	
20	Vor	Nach der Parametrierung **ist** die Verbindung zwischen dem Regler und dem PC **zu trennen**.	**#4**
	Nach	Nach der Parametrierung **trennen Sie** die Verbindung zwischen dem Regler und dem PC.	
21	Vor	Sind mehrere Geräte im Netzwerk vorhanden, so **ist** die Adresse des gewünschten Gerätes **auszuwählen**.	**#4**
	Nach	Sind mehrere Geräte im Netzwerk vorhanden, **wählen Sie** die Adresse des gewünschten Gerätes **aus**.	
22	Vor	Um den Sollwert zu erreichen, **ist** die Konfigurierung des Anschlusses **zu berücksichtigen**.	**#4**
	Nach	Um den Sollwert zu erreichen, **berücksichtigen Sie** die Konfigurierung des Anschlusses.	
23	Vor	Zum Anschluss an den PC **sind** die beiliegenden Kabel miteinander **zu verbinden**.	**#4**
	Nach	Zum Anschluss an den PC **verbinden Sie** die beiliegenden Kabel miteinander.	

24	Vor	Gerät und Netzkabel **sind** von Kindern unter 8 Jahren **fernzuhalten**.	**#6**
	Nach	**Halten Sie** das Gerät und das Netzkabel von Kindern unter 8 Jahren **fern**.	

Regel 8 – Überflüssige Präfixe vermeiden

	KS	Sätze	Ref. Quelle
1	Vor	Das Tool **bietet** diese Korrekturen bei der Eingabe der Werte im Bereich Übersicht **an**.	**#4**
	Nach	Das Tool **bietet** diese Korrekturen bei der Eingabe der Werte im Bereich Übersicht.	
2	Vor	Sie können im nächsten Schritt ein Installationsverzeichnis für das Tool **auswählen**.	**#4**
	Nach	Sie können im nächsten Schritt ein Installationsverzeichnis für das Tool **wählen**.	
3	Vor	Bevor Sie das Gerät in Betrieb nehmen, **lesen** Sie zuerst die Bedienungsanleitung aufmerksam **durch**.	**#8**
	Nach	Bevor Sie das Gerät in Betrieb nehmen, **lesen** Sie zuerst die Bedienungsanleitung aufmerksam.	
4	Vor	**Wählt** man einen bestimmten Zeichensatz als Standardwert **aus**, wird dieser Zeichensatz in allen Stationen verwendet.	**#4**
	Nach	**Wählt** man einen bestimmten Zeichensatz als Standardwert, wird dieser Zeichensatz in allen Stationen verwendet.	
5	Vor	**Kaufen** Sie die Geräte in einem anderen Land **ein**, werden Garantieleistungen nur in diesem Land erbracht.	**#6**
	Nach	**Kaufen** Sie die Geräte in einem anderen Land, werden Garantieleistungen nur in diesem Land erbracht.	

6	Vor	Die Kleberreste **haften** auf der Oberfläche **an**, wenn sie nicht schnell entfernt werden.	**#1**
	Nach	Die Kleberreste **haften** auf der Oberfläche, wenn sie nicht schnell entfernt werden.	
7	Vor	**Überprüfen** Sie die Adresse des Ports im Geräte-Manager.	**#4**
	Nach	**Prüfen** Sie die Adresse des Ports im Geräte-Manager.	
8	Vor	Durch Anklicken des Buttons Zusatzinformation **speichert** das System die Angaben **ab**.	**#4**
	Nach	Durch Anklicken des Buttons Zusatzinformation **speichert** das System die Angaben.	
9	Vor	Der Spediteur **liefert** die Ware zum vereinbarten Termin **an**.	**#3**
	Nach	Der Spediteur **liefert** die Ware zum vereinbarten Termin.	
10	Vor	**Wählen** Sie die Option "Software von einer bestimmten Liste installieren" **aus**.	**#4**
	Nach	**Wählen** Sie die Option "Software von einer bestimmten Liste installieren".	
11	Vor	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse **ein**.	**#8**
	Nach	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse.	
12	Vor	Sie können die angezeigten Werte lokal auf der Festplatte **abspeichern**.	**#4**
	Nach	Sie können die angezeigten Werte lokal auf der Festplatte **speichern**.	
13	Vor	Sorgen Sie dafür, dass die Quelle ein einwandfreies Signal **absendet**.	**#8**
	Nach	Sorgen Sie dafür, dass die Quelle ein einwandfreies Signal **sendet**.	
14	Vor	**Schicken** Sie das Gerät zusammen mit dem Original-Kaufbeleg an nachstehende Adresse **zu**.	**#9**
	Nach	**Schicken** Sie das Gerät zusammen mit dem Original-Kaufbeleg an nachstehende Adresse.	

15	Vor	Wir **senden** Ihnen innerhalb 24 Stunden einen Paketaufkleber für die kostenlose Rücksendung **zu**.	**#9**
	Nach	Wir **senden** Ihnen innerhalb 24 Stunden einen Paketaufkleber für die kostenlose Rücksendung.	
16	Vor	**Überprüfen** Sie, ob ausreichend Wasser im Wassertank vorhanden ist.	**#1**
	Nach	**Prüfen** Sie, ob ausreichend Wasser im Wassertank vorhanden ist.	
17	Vor	In der zweiten Phase **sendet** die Quelle ein Signal **ab**.	**#8**
	Nach	In der zweiten Phase **sendet** die Quelle ein Signal.	
18	Vor	**Speichern** Sie die angezeigten Werte lokal auf der Festplatte **ab**.	**#4**
	Nach	**Speichern** Sie die angezeigten Werte lokal auf der Festplatte.	
19	Vor	Falls Ihr Receiver diese Möglichkeiten nicht **anbietet**, können Sie nur die wichtigsten Einstellungen vornehmen.	**#8**
	Nach	Falls Ihr Receiver diese Möglichkeiten nicht **bietet**, können Sie nur die wichtigsten Einstellungen vornehmen.	
20	Vor	Wenn Sie die Geräte in einem anderen Land **einkaufen**, werden Garantieleistungen nur in diesem Land erbracht.	**#6**
	Nach	Wenn Sie die Geräte in einem anderen Land **kaufen**, werden Garantieleistungen nur in diesem Land erbracht.	
21	Vor	Bevor Sie das Gerät in Betrieb nehmen, müssen Sie zuerst die Bedienungsanleitung aufmerksam **durchlesen**.	**#8**
	Nach	Bevor Sie das Gerät in Betrieb nehmen, müssen Sie zuerst die Bedienungsanleitung aufmerksam **lesen**.	
22	Vor	Sofern auf der Oberfläche Kleberreste **anhaften**, sind diese vollständig zu entfernen.	**#1**
	Nach	Sofern auf der Oberfläche Kleberreste **haften**, sind diese vollständig zu entfernen.	
23	Vor	Tab 7 und Tab 8 **zeigen** die verfügbaren Anwendungen **an**.	**#4**

	Nach	Tab 7 und Tab 8 **zeigen** die verfügbaren Anwendungen.	
24	Vor	In Tab 7 und Tab 8 **werden** die verfügbaren Anwendungen **angezeigt**.	**#4**
	Nach	In Tab 7 und Tab 8 **werden** die verfügbaren Anwendungen **gezeigt**.	

Regel 9 – Keine Wortteile weglassen

	KS	Sätze	Ref. Quelle
1	Vor	**Elektro- und Gasgeräte** dürfen nur von geschulten Fachleuten installiert werden.	**#7**
	Nach	**Elektrogeräte und Gasgeräte** dürfen nur von geschulten Fachleuten installiert werden.	
2	Vor	Innerhalb der Garantiezeit beseitigen wir alle Mängel des Gerätes, die auf **Material- oder Fabrikationsfehlern** beruhen.	**#6**
	Nach	Innerhalb der Garantiezeit beseitigen wir alle Mängel des Gerätes, die auf **Materialfehlern oder Fabrikationsfehlern** beruhen.	
3	Vor	Sogar **Soja- und laktosefreie Milch** lassen sich mit dieser Maschine perfekt aufschäumen.	**#6**
	Nach	Sogar **Sojamilch und laktosefreie Milch** lassen sich mit dieser Maschine perfekt aufschäumen.	
4	Vor	Das Gleiche gilt bei Nichtbeachtung der **Gebrauchs-, Pflege- und Wartungsanweisung**.	**#6**
	Nach	Das Gleiche gilt bei Nichtbeachtung der **Gebrauchsanweisung, Pflegeanweisung und Wartungsanweisung**.	
5	Vor	Die Teppichböden sind entsprechend den **Liefer- und Zahlungsbedingungen** zu prüfen.	**#1**
	Nach	Die Teppichböden sind entsprechend den **Lieferbedingungen und Zahlungsbedingungen** zu prüfen.	

6	Vor	Hartnäckige Flecken, wie Fettspritzer, **Lack- oder Klebstoffreste**, sind mit handelsüblichem Kunststoffreiniger eventuell zu beseitigen.	**#7**
	Nach	Hartnäckige Flecken, wie Fettspritzer, **Lackreste oder Klebstoffreste**, sind mit handelsüblichem Kunststoffreiniger eventuell zu beseitigen.	
7	Vor	Wir raten Ihnen, die **Bedienungs- und Pflegehinweise** des Herstellers genauestens zu lesen.	**#7**
	Nach	Wir raten Ihnen, die **Bedienungshinweise und Pflegehinweise** des Herstellers genauestens zu lesen.	
8	Vor	Dies ist nicht auf einen **Konstruktions- oder Verarbeitungsfehler** in unseren Möbeln zurückzuführen.	**#7**
	Nach	Dies ist nicht auf einen **Konstruktionsfehler oder Verarbeitungsfehler** in unseren Möbeln zurückzuführen.	
9	Vor	Die wichtigsten Parameter der **Ein- und Ausgangskonfiguration** sind voreingestellt.	**#4**
	Nach	Die wichtigsten Parameter der **Eingangskonfiguration und Ausgangskonfiguration** sind voreingestellt.	
10	Vor	Im Falle der Auswahl der freien Konfiguration kann der **Start- und Endpunkt** frei gewählt werden.	**#4**
	Nach	Im Falle der Auswahl der freien Konfiguration können der **Startpunkt und der Endpunkt** frei gewählt werden.	
11	Vor	Die **Ist- und Sollwerte** des zweiten Regelkreises werden nach der Konfiguration angezeigt.	**#4**
	Nach	Der **Istwert und der Sollwert** des zweiten Regelkreises werden nach der Konfiguration angezeigt.	
12	Vor	Schützen Sie das Gerät vor **Tropf- und Spritzwasser**.	**#8**
	Nach	Schützen Sie das Gerät vor **Tropfwasser und Spritzwasser**.	
13	Vor	**Kalk- und Wasserflecken** beseitigen Sie mit dem vom Hersteller empfohlenen Spezialreiniger.	**#7**
	Nach	**Kalkflecken und Wasserflecken** beseitigen Sie mit dem vom Hersteller empfohlenen Spezialreiniger.	

14	Vor	Es ist darauf zu achten, dass die **Garn- oder Mikrofaserpads** regelmäßig gewechselt werden.	**#1**
	Nach	Es ist darauf zu achten, dass die **Garnpads oder Mikrofaserpads** regelmäßig gewechselt werden.	
15	Vor	**Kunststoffgriffe und -deckelknöpfe** werden bei Verwendung im Backofen heiß.	**#2**
	Nach	**Kunststoffgriffe und Kunststoffdeckelknöpfe** werden bei Verwendung im Backofen heiß.	
16	Vor	Trennen Sie alle **Spannungs- und Druckquellen** von der Maschine.	**#3**
	Nach	Trennen Sie alle **Spannungsquellen und Druckquellen** von der Maschine.	
17	Vor	Eine Bevorratung der wichtigsten **Ersatz- und Verschleißteile** ist eine wichtige Voraussetzung für ständige Einsatzbereitschaft.	**#3**
	Nach	Eine Bevorratung der wichtigsten **Ersatzteile und Verschleißteile** ist eine wichtige Voraussetzung für ständige Einsatzbereitschaft.	
18	Vor	Wechseln der Schneidköpfe ist **prozess- und produktabhängig**.	**#3**
	Nach	Wechseln der Schneidköpfe ist **prozessabhängig und produktabhängig**.	
19	Vor	Die **Sicherheits- und Anwendungsrichtlinien** der Reinigungsmittelhersteller sind stets einzuhalten.	**#1**
	Nach	Die **Sicherheitsrichtlinien und Anwendungsrichtlinien** der Reinigungsmittelhersteller sind stets einzuhalten.	
20	Vor	In **Eingangs- und Übergangsbereichen** sind große Eingangsmatten fest zu installieren.	**#1**
	Nach	In **Eingangsbereichen und Übergangsbereichen** sind große Eingangsmatten fest zu installieren.	
21	Vor	Die **Anwärm- und Entwässerungsvorgänge** sind gemäß der Betriebsanleitung zu beachten.	**#3**
	Nach	Der **Anwärmvorgang und der Entwässerungsvorgang** sind gemäß der Betriebsanleitung zu beachten.	

22	Vor	Die **Reparatur- und Wartungsarbeiten** der Maschine sind ausschließlich von einem Industriemechaniker durchzuführen.	#3
	Nach	Die **Reparaturarbeiten und Wartungsarbeiten** der Maschine sind ausschließlich von einem Industriemechaniker durchzuführen.	
23	Vor	Die Seite I/O Konfigurierung steht zur detaillierten Parametrierung der **Ein- und Ausgänge** zur Verfügung.	#4
	Nach	Die Seite I/O Konfigurierung steht zur detaillierten Parametrierung der **Eingänge und Ausgänge** zur Verfügung.	
24	Vor	Prüfen Sie, ob sich **Wasser-**, **Gasrohre** oder stromführende Leitungen im Bohrbereich befinden.	#7
	Nach	Prüfen Sie, ob sich **Wasserrohre**, **Gasrohre** oder stromführende Leitungen im Bohrbereich befinden.	

References

ABUEUS – Arbeitsbereich Usability-Engineering der Universität des Saarlandes. 2006. *Methoden und Verfahren: Eyetracking*. http://usability.is.uni-sb.de/methoden/eyetracking.php.

Adams, Tim. 2010. Can Google break the computer language barrier? *The Observer*. https://www.theguardian.com/technology/2010/dec/19/google-translate-computers-languages.

Aikawa, Takako, Lee Schwartz, Ronit King, Mo Corston-Oliver & Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In Bente Maegaard (Hrsg.), *Proceedings of Machine Translation Summit XI*, 1–7. Copenhagen: European Association for Machine Translation. https://aclanthology.org/2007.mtsummit-papers.1.

Alabau, Vicent, Ragnar Bonkb, Christian Buckc, Michael Carlb, Francisco Casacubertaa, Mercedes García-Martínezb, Jesús Gonzálezа, Philipp Koehnc, Luis Leivaa, Bartolomé Mesa-Laob, Daniel Ortiza, Herve Saint-Amandc, Germán Sanchisa & Chara Tsoukalac. 2013. ASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics* 100(1). 101–112.

Allen, Jeffrey. 2003. Post-editing. In Harold Somers (Hrsg.), *Computers and translation: A translator's guide* (Benjamins Translation Library 35), 297–318. Amsterdam: Benjamins. DOI: 10.1075/btl.35.19all.

ALPAC. 1966. *Language and machines: Computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.

Al-Ansary, Sameh. 2011. Interlingua-based machine translation systems: UNL versus other interlinguas. In *Proceedings of the 11th international conference on language engineering*. Kairo. https://www.bibalex.org/isis/UploadedFiles/Publications/Cairo2011a_1.pdf.

Aranberri, Nora & Johann Roturier. 2009. Comparison of alternatives to strict source control: A case study with –ing words. In *Pre-Proceedings of the Workshop on Controlled Natural Language*. Marettimo Island.

Arnold, Doug. 1994. *Machine translation: An introductory guide*. Oxford: NCC Blackwell.

Attila, Görög. 2014. Quality evaluation today: The dynamic quality framework. *Translating and the Computer* 36. http://www.mt-archive.info/10/Asling-2014-Gorog.pdf.

Avramidis, Eleftherios, Aljoscha Burchardt, Christian Federmann, Maja Popovic, Cindy Tscherwinka & David Vilar. 2014. Involving language professionals in the evaluation of machine translation. *Language Resources and Evaluation* 48(4). 541–559.

Avramidis, Eleftherios & Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL), Human Language Technologies*. Columbus, OH: 763–770.

Aziz, Wilker, Sheila C. M. de Sousa & Lucia Specia. 2012. PET: A tool for post-editing and assessing machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3982–3987. Istanbul: European Language Resources Association (ELRA). https://aclanthology.org/L12-1587/.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio & Yann LeCun (Hrsg.), *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA: arXiv. DOI: 10.48550/arXiv.1409.0473.

Balkan, Lorna & Frederik Fouvry. 1995. *Corpus-based test suite generation*. Report to LRE 62-089 (D-WP5.2), Test Suites for Natural Language Processing (TSNLP). University of Essex.

Balkan, Lorna, Siety Meijer, Doug Arnold, Eva Dauphin, Dominique Estival, Kirsten Falkedal, Sabine Lehmann, Klaus Netter & Sylvie Regnier-Prost. 1994. *Issues in test suite design*. Report to LRE 62-089 (D-WP2.1). University of Essex.

Banerjee, Satanjeev & Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin & Clare Voss (Hrsg.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 65–72. Ann Arbor, MI: Association for Computational Linguistics. https://aclanthology.org/W05-0909.

Baumert, Andreas & Annette Verhein-Jarren. 2012. *Texten für die Technik: Leitfaden für Praxis und Studium*. Berlin: Springer.

Beeby, Allison, Mònica Fernández, Olivia Fox, Amparo Hurtado Albir, Anna Kuznik, Wilhelm Neunzig, Patricia Rodríguez Inés, Lupe Romero & Stefanie Wimmer. 2011. Results of the validation of the PACTE translation competence model: Translation problems and translation competence. In Cecilia Alvstad, Adelina Hild & Elisabet Tiselius (Hrsg.), *Methods and strategies of process research: Integrative approaches in translation studies* (Benjamins Translation Library 94), 317–343. Amsterdam: Benjamins. DOI: 10.1075/btl.94.22pac.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In Jian Su, Kevin Duh & Xavier Carreras (Hrsg.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–267. Austin, TX. DOI: 10.18653/v1/d16-1025.

Bernth, Arendse. 1999. Controlling input and output of MT for greater acceptance. In *Proceedings of the 21st ASLIB Conference, Translating and the Computer*. London: Aslib. https://aclanthology.org/1999.tc-1.13.

Bernth, Arendse & Claudia Gdaniec. 2001. MTranslatability. *Machine Translation* 16(3). 175–218.

Beyer, Anne, Vivien Macketanz, Aljoscha Burchardt & Philip Williams. 2017. Can out-of-the-box NMT beat a domain-trained moses on technical data? In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT): User Studies and Project/Product Descriptions*, 41–46. Prague. https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/Conference_Booklet_EAMT2017.pdf.

Bojar, Ondřej. 2011. Analysing error types in English-Czech machine translation. *Bull Math Linguist* 95. 63–76.

Bredel, Ursula & Christiane Maaß. 2016. *Duden: Leichte Sprache. Theoretische Grundlagen, Anleitung für die Praxis*. Berlin: Bibliographisches Institut.

Bruckner, Christine. 2020. Maschinelle Übersetzung: Wer tut was im Markt? *MDÜ Fachzeitschrift für Dolmetscher und Übersetzer* 1/2020. 44–48.

Brunette, Louise. 2000. Towards a terminology for translation quality assessment. *The Translator* 6(2). 169–182.

Burchardt, Aljoscha & Kim Harris. 2017. Improving machine translation: The gap between research approaches and industry needs. In Jörg Porsiel (Hrsg.), *Maschinelle Übersetzung: Grundlagen für den professionellen Einsatz*, 126–139. Berlin: BDÜ Fachverlag.

Burchardt, Aljoscha, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter & Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *The Prague Bulletin of Mathematical Linguistics* 108(1). 159–170.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz & Josh Schroeder. 2007. (Meta-)Evaluation of machine translation. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce & Christof Monz (Hrsg.), *Proceedings of the Second Workshop on Statistical Machine Translation*, 136–158. Prague: Association for Computational Linguistics. https://www.aclweb.org/anthology/W07-0718 (28 Januar, 2017).

Callison-Burch, Chris, Philipp Koehn, Christof Monz & Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, 1–28. Athens. https://aclanthology.org/W09-0401.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia (Hrsg.), *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 10–51. Montréal: Association for Computational Linguistics. https://www.aclweb.org/anthology/W12-3102.

Campbell, Stuart. 2000. Critical structures in the evaluation of translations from Arabic into English as a second language. *The Translator* 6(2). 211–229.

Canfora, Carmen & Angelika Ottmann. 2015. Risikomanagement für Übersetzungen. *trans-kom* 8(2). 314–346.

Caplin, Andrew & Paul W. Glimcher. 2014. *Diminishing marginal utility: an overview*. www.sciencedirect.com.

Carl, Michael. 2012. Translog – II: A program for recording user activity data for empirical reading and writing research. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Eight International Conference on Language Resources and Evaluation*, 4108–4112. Istanbul: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/614_Paper.pdf.

Carl, Michael, Silke Gutermuth & Silvia Hansen-Schirra. 2015. Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings. In Aline Ferreira & John W. Schwieter (Hrsg.), *Psycholinguistic and cognitive inquiries into translation and interpreting*, 145–174. Amsterdam: Benjamins.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer & Ralf Klabunde. 2010. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. 3. Überarb. u. erw. Aufl. Heidelberg: Spektrum, Akademischer Verlag.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer. 2004. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. 2. Überarb u. erw Aufl. Heidelberg: Spektrum, Akademischer Verlag.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley & Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108(1). 109–120.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Miceli Valerio, Antoni Barone & Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. In Sadao Kurohashi & Pascale Fung (Hrsg.), *Proceedings of Machine Translation Summit 2017*, Bd. 1: Research Track, 116–131. Nagoya: Asia-Pacific Association for Machine Translation (AAMT). https://www.computing.dcu.ie/~away/PUBS/2017/SheilaMTSummit2017.pdf.

Castilho, Sheila & Sharon O'Brien. 2016. Content profiling and translation scenarios. *The Journal of Internationalization and Localization* 3(1). 18–37. DOI: 10.1075/jial.3.1.02cas.

Cavalli-Sforza, Violetta & Alon Lavie. 2006. *Hybrid machine translation: Why and how?* Cambridge, MA. https://aclanthology.org/2006.amta-panels.0 (26 Oktober, 2017).

Chatterjee, Rajen, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia & Frederic Blain. 2017. Guiding neural machine translation decoding with external knowledge. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn & Julia Kreutzer (Hrsg.), *Proceedings of the Conference on Machine Translation (WMT)*, 157–168. Copenhagen: Association for Computational Linguistics. DOI: 10.18653/v1/W17-4716.

Chen, Boxing, Roland Kuhn & Samuel Larkin. 2012. Port: a precision-order-recall MT evaluation metric for tuning. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee & Jong C. Park (Hrsg.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Bd. 1: Long Papers, 930–939. Jeju Island: Association for Computational Linguistics. https://aclanthology.org/P12-1098.

Chernick, Michael R. 2008. *Bootstrap methods: A guide for practioners and researchers*. 2. Aufl. Hoboken, NJ: Wiley.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang & Walter Daelemans (Hrsg.), *Proceedings*

of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734. Doha: Association for Computational Linguistics. DOI: 10.3115/v1/D14-1179.

Condon, Sherri, Dan Parvaz, John Aberdeen, Christy Doran, Andrew Freeman & Marwan Awad. 2010. Evaluation of machine translation errors in English and Iraqi Arabic. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta: European Language Resources Association (ELRA). https://aclanthology.org/L10-1065/.

Congree. 2018. Congree-Stilregel. Eine nicht-veröffentlichte Publikation vom 21.11.2018.

Constant, Mathieu, Gülsen Eryiğit, Johanna Monti, Lonneke van Der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43(4). 837–892.

Correa, Nelson. 2003. A fine-grained evaluation framework for machine translation system development. In *Proceedings of Machine Translation Summit IX*. New Orleans, LA: Association for Machine Translation in the Americas. https://aclanthology.org/2003.mtsummit-papers.7.

Costa, Angela, Wang Ling, Tiago Luís, Rui Correia & Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation* 29. 127–161. DOI: 10.1007/s10590-015-9169-0.

Costa-Jussà, Marta R. & José A. R. Fonollosa. 2015. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language* 32(1). 3–10.

Coughlin, Deborah. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of Machine Translation Summit IX*, 63–70. New Orleans, LA: Association for Machine Translation in the Americas. https://aclanthology.org/2003.mtsummit-papers.9/.

Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou & Peter Zoldan. 2016. *SYSTRAN's pure neural machine translation systems*. DOI: 10.48550/arXiv.1610.05540.

Creswell, John W. & Vicki L. Plano Clark. 2007. *Designing & conducting mixed methods research + the mixed methods reader (bundle)*. London: Sage.

DeepL. 2020. *Passen Sie den DeepL Übersetzer mit Ihrem individuellen Glossar an.* https://www.deepl.com/de/blog/20200506.html (30 Juni, 2020).
Denkowski, Michael & Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA): Research Papers*. Denver, CO: Association for Machine Translation in the Americas. https://aclanthology.org/2010.amta-papers.20.
Denkowski, Michael & Alon Lavie. 2012. Challenges in predicting machine translation utility for human post-editors. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas (AMTA): Research Papers*. San Diego, CA: Association for Machine Translation in the Americas. https://aclanthology.org/2012.amta-papers.6.
Dinu, Georgiana, Prashant Mathur, Marcello Federico & Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In Anna Korhonen, David Traum & Lluís Màrquez (Hrsg.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3063–3068. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/P19-1294.
Doddington, George. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of 2nd Human Language Technologies Conference (HLT-02)*, 138–145. San Diego, CA: Morgan Kaufmann. https://dl.acm.org/doi/10.5555/1289189.1289273.
Doherty, Stephen. 2012. *Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-method approach*. Dublin: Dublin City University. (Diss.).
Doherty, Stephen. 2017. Issues in human and automatic translation quality assessment. In Dorothy Kenny (Hrsg.), *Human issues in translation technology*, 131–186. London: Routledge.
Doherty, Stephen & Sharon O'Brien. 2009. Can MT output be evaluated through eye tracking? In *Proceedings of Machine Translation Summit XII: Posters*, 214–221. Ottawa: Association for Machine Translation in the Americas (AMTA). https://aclanthology.org/2009.mtsummit-posters.5/.
Doherty, Stephen & Sharon O'Brien. 2012. A user-based usability assessment of raw machine translated technical instructions. In *Conference of the Association for Machine Translation in the Americas*. San Diego, CA: Association for Machine Translation in the Americas (AMTA). https://aclanthology.org/2012.amta-commercial.4.

Dorr, Bonnie, Pamela W. Jordan & John W. Benoit. 1999. A survey of current paradigms in machine translation. *Advances in Computers* 49(2). 1–68.

Douglas, Shona & Matthew Hurst. 1996. Controlled language support for Perkins Approved Clear English (PACE). In *Proceedings of the First International Workshop on Controlled Language Applications (CLAW)*, 93–105. Leuven: Katholieke Universiteit Leuven. https://aclanthology.org/www.mt-archive.info/90/CLAW-1996-Douglas.pdf.

Drewer, Petra & Klaus-Dirk Schmitz. 2017. *Terminologiemanagement: Grundlagen – Methoden – Werkzeuge* (Kommunikation und Medienmanagement). Berlin: Springer.

Drewer, Petra & Wolfgang Ziegler. 2014. *Technische Dokumentation: Übersetzungsgerechte Texterstellung und Content-Management*. 2. Aufl. Würzburg: Vogel.

Drugan, Joanna. 2013. *Quality in professional translation: Assessment and improvement*. London: Bloomsbury.

DuBay, William H. 2004. *The principles of readability*. Costa Mesa, CA: Impact Information. https://files.eric.ed.gov/fulltext/ED490073.pdf.

Dubey, Shantanoo. 2017. Survey of machine translation techniques. *International Journal of Advance Research in Computer Science and Management Studies* 5(2).

Duchowski, Andrew. 2007. *Eye tracking methodology: Theory and practice*. London: Springer.

Dyson, Mary & Jean Hannah. 1987. Toward a methodology for the evaluation of machine assisted translation systems. *Computers and Translation* 2(3). 163–176.

Eckstein, Peter P. 2008. *Angewandte Statistik mit SPSS: praktische Einführung für Wirtschaftswissenschaftler*. Wiesbaden: Gabler.

Eisele, Andreas. 2007. Hybrid machine translation: combining rule-based and statistical MT systems. In *Conference of the First Machine Translation Marathon*. Edinburgh.

Eisele, Andreas, Christian Federmann, Hans Uszkoreit, Hervé Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann & Yu Chen. 2008. Hybrid architectures for multi-engine machine translation. In *Proceedings of Translating and the Computer 30*. London: Aslib. https://aclanthology.org/2008.tc-1.2/.

Eisold, Christian. 2017. Zur Rolle der Terminologie in der maschinellen Übersetzung. In Jörg Porsiel (Hrsg.), *Maschinelle Übersetzung: Grundlagen für den professionellen Einsatz*, 109–125. Berlin: BDÜ-Fachverlag.

Elliott, Debbie, Anthony Hartley & Eric Atwell. 2004. A fluency error categorization scheme to guide automated machine translation evaluation. In Robert E. Frederking & Kathryn B. Taylor (Hrsg.), *Machine translation: From real users to*

research: 6th conference of the Association for Machine Translation in the Americas (AMTA 2004), Washington, DC (Lecture Notes in Computer Science 3265), 64–73. Berlin: Springer.

Elliston, John S. G. 1979. Computer-aided translation: a business viewpoint. In Barbara M. Snell (Hrsg.), *Proceedings of Translating and the Computer 1*, 149–158. London: Aslib. https://aclanthology.org/1978.tc-1.8.

Engelberg, Stefan. 2009. *Korpuslinguistik*. http : / / www1 . ids - mannheim . de / fileadmin/lexik/lehre/engelberg/Webseite_MethLex/Korpuslinguistik-V1.pdf (20 September, 2017).

Erzberger, Christian & Udo Kelle. 2003. Making inferences in mixed methods: The rules of integration. In Abbas Tashakkori & Charles Teddlie (Hrsg.), *Handbook of mixed methods for the social & behavioral sciences*, 457–490. Thousand Oaks, CA: Sage.

Fabienne, Cap. 2016. *Introduction to neural machine translation*. https://cl.lingfil.uu.se/kurs/MT18/slides/f7-nmt1.pdf (31 Juli, 2018).

Farkas, David. 1985. The concept of consistency in writing and editing. *Journal of Technical Writing and Communication* 15(4). 353–364.

Farrús, Mireia, Marta Ruiz Costa-jussa, Jose Bernardo Mariño Acebal & José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In François Yvon & Viggo Hansen (Hrsg.), *Proceedings of the 14th Annual conference of the European Association for Machine Translation (EAMT)*. Saint Raphaël: European Association for Machine Translation. https://aclanthology.org/2010.eamt-1.12.

Fettke, Peter. 2016. *Client-Server-Architektur*. http : / / www . enzyklopaedie - der - wirtschaftsinformatik . de / lexikon / is - management / Systementwicklung / Softwarearchitektur / Architekturparadigmen / Client - Server - Architektur (20 Januar, 2017).

Fiederer, Rebecca & Sharon O'Brien. 2009. Quality and machine translation: A realistic objective? *The Journal of Specialised Translation* 11. 52–74.

Fishel, Mark, Ondřej Bojar & Maja Popović. 2012. Terra: A collection of translation error-annotated corpora. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 7–14. Istanbul: European Language Resources Association (ELRA). https://aclanthology.org/L12-1260/.

Flanagan, Mary. 1994. Error classification for MT evaluation. In *Proceedings of the Association of Machine Translation of the Americas*, 65–72. Washington

D.C.: Association of Machine Translation of the Americas (AMTA). https://aclanthology.org/1994.amta-1.9.

Fleury, Isabelle. o. D. *tekom veröffentlicht Leitlinie „Regelbasiertes Schreiben: Deutsch für die Technische Kommunikation“*. https://www.dokuworld.de/newsreader/tekom-veroeffentlicht-leitlinie-regelbasiertes-schreiben-deutsch-fuer-die-technische-kommunikation.html.

Forcada, Mikel L. 2010. Machine translation today. In Yves Gambier & Luc van Doorslaer (Hrsg.), *Handbook of translation studies*, Bd. 1, 215–223. Amsterdam: Benjamins.

Frey, Lawrence R., Carl H. Botan & Gary L. Kreps. 1991. *Investigating communication: An introduction to research methods*. Englewood Cliffs, NJ: Prentice Hall.

Funke, Joachim. 2006. Wenn Blicke sprechen. *Ruperto Carola* (1). http://www.uni-heidelberg.de/presse/ruca/ruca06-1/wenn.html.

Gamallo, Pablo & Marcos Garcia. 2019. Unsupervised compositional translation of multiword expressions. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (Hrsg.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 40–48. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5106.

Geldbach, Stefanie. 2009. Neue Werkzeuge zur Autorenunterstützung: Quelltextbearbeitung in Kombination mit Translation-Memory-Systemen. *MDÜ: Fachzeitschrift für Dolmetscher und Übersetzer* (4). 10–19.

Gerlach, Johanna. 2015. *Improving statistical machine translation of informal language: a rule-based pre-editing approach for French forums*. Genf: Universität Genf. (Diss.).

Gerlach, Johanna, Victoria Porro Rodriguez, Pierrette Bouillon & Sabine Lehmann. 2013. Combining pre-editing and post-editing to improve SMT of user-generated content. In Sharon O'Brien, Michel Simard & Lucia Specia (Hrsg.), *Proceedings of Machine Translation Summit XIV Workshop on Post-editing Technology and Practice*, 45–53. Nice. https://aclanthology.org/2013.mtsummit-wptp.6.

Gesellschaft für Technische Kommunikation – tekom e. V. 2009. *Arbeitsgruppe Technisches Deutsch*. http://www.tekom.de/index_neu.jsp?url=/servlet/ControllerGUI?action=voll&id=2742 (6 Dezember, 2014).

Gesellschaft für Technische Kommunikation – tekom e.V. 2013. *Leitlinie „Regelbasiertes Schreiben, Deutsch für die Technische Kommunikation"*. 2. Erweiterte Auflage. Stuttgart.

Giménez, Jesús & Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder & Cameron Shaw Fordyce (Hrsg.), *Proceedings of the Third Work-*

shop on Statistical Machine Translation, 195–198. Columbus, OH: Association for Computational Linguistics. https://aclanthology.org/W08-0332.

Gonzàlez, Meritxell & Jesús Giménez. 2014. *An open toolkit for automatic machine translation (meta-)evaluation.* Techn. Ber. LSI-14-2-T. Technical Manual v3.0. February 2014. Departamento de Lenguajes y Sistemas Informáticos, Universitat Politècnica de Catalunya. http://asiya.lsi.upc.edu/Asiya_technical_manual_v3.0.pdf.

Göpferich, Susanne. 2001. Von Hamburg nach Karlsruhe: Ein kommunikationsorientierter Bezugsrahmen zur Bewertung der Verständlichkeit von Texten. *Fachsprache* 23(3–4). 117–138.

Göpferich, Susanne. 2007a. Sprachstandard oder Kontrollmechanismus? Textqualität steuern mit kontrollierter Sprache. *Technische Kommunikation Fachzeitschrift für technische Dokumentation und Informationsmanagement* 4. 16–21.

Göpferich, Susanne. 2007b. Standardisierung von Kommunikation. In Karlfried Knapp, Gerd Antos, Michael Becker-Mrotzek, Arnulf Deppermann, Susanne Göpferich, Joachim Grabowski, Michael Klemm & Claudia Villiger (Hrsg.), *Angewandte Lingistik*, 479–502. Tübingen; Basel: A. Francke Verlag.

Göpferich, Susanne. 2008. *Textproduktion im Zeitalter der Globalisierung: Entwicklung einer Didaktik des Wissenstransfers.* 3. Aufl. Tübingen: Stauffenburg Verlag.

Goshawke, Walter, Ian D. K. Kelly & J. David Wigg. 1987. *Computer translation of natural language.* Wilmslow: Sigma Press.

Govyaerts, Patrick. 1996. Controlled English, curse or blessing? A user's perspective. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996)*, 137–142. Leuven: Centre for Computational Linguistics.

Grice, Paul. 1975. Logic and conversation. In Paul Grice (Hrsg.), *Studies in the way of words 1989*, 22–40. Harvard: Harvard University press. https://courses.media.mit.edu/2004spring/mas966/Grice%20Logic%20and%20Conversation.pdf.

Groeben, Norbert. 1982. *Leserpsychologie: Textverständnis – Textverständlichkeit.* Münster: Aschendorff.

Groves, Declan. 2007. *Hybrid data-driven models of machine translation.* Dublin City University. (Diss.).

Guillou, Liane & Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC 2016), 636–643. Paris: European Language Resources Association (ELRA). https://aclanthology.org/L16-1100.

Guzmán, Francisco, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad & Stephan Vogel. 2015. How do humans evaluate machine translation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva & Pavel Pecina (Hrsg.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 457–466. Lisbon: Association for Computational Linguistics. DOI: 10.18653/v1/W15-3059.

Haller, Johann & Jörg Schütz. 2001. CLAT: Controlled language authoring technology. In Mary J. Northrop & Scott Tilley (Hrsg.), *Proceedings of the 19th Annual International Conference of Computer Documentation*, 78–82. New York: ACM Press.

Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hamon, Olivier. 2007. Assessing human and automated quality judgments in the French MT-evaluation campaign CESTA. In Bente Maegaard (Hrsg.), *Proceedings of Machine Translation Summit XI*, 231–238. Copenhagen. https://aclanthology.org/2007.mtsummit-papers.31.

Han, Aaron L. F. 2018. Machine translation evaluation resources and methods: A survey. In *IPRC-2018 (Ireland Postgraduate Research Conference)*. arXiv. https://arxiv.org/pdf/1605.04515 (17 Oktober, 2018).

Han, Aaron L. F., Derek F. Wong & Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In Martin Kay & Christian Boitet (Hrsg.), *Proceedings of COLING: Posters*, 441–450. Mumbai: The COLING 2012 Organizing Committee. https://aclanthology.org/C12-2044.

Han, Aaron L. F., Derek F. Wong & Lidia S. Chao. 2017. *Machine translation evaluation resources and methods: A survey*. https://arxiv.org/pdf/1605.04515 (20 Dezember, 2017).

Han, Aaron L. F., Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing & Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In Andy Way, Khalil Sima'an & Mikel L. Forcada (Hrsg.), *Proceedings of the XIV Machine Translation Summit: Posters*, 215–222. Nice. https://aclanthology.org/2013.mtsummit-posters.3.

Hannu, Kuusela & Paul Pallab. 2000. A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology* 113(3). 387–404.

Hansen-Schirra, Silvia & Silke Gutermuth. 2018. Modellierung und Messung einfacher und leichter Sprache. In Jekat Susanne, Martin Kappus & Klaus Schubert (Hrsg.), *Barrieren abbauen, Sprache gestalten* (Working Papers in Applied

Linguistics), 7–23. Winterthur: ZHAW Zürcher Hochschule für Angewandte Wissenschaften.
Hansen-Schirra, Silvia & Christiane Maaß (Hrsg.). 2020. *Easy language research: Text and user perspectives*. Berlin: Frank & Timme.
Hansen-Schirra, Silvia, Moritz Schäffer & Jean Nitzke. 2017. Post-Editing: Strategien, Qualität, Effizienz. In Jörg Porsiel (Hrsg.), *Maschinelle Übersetzung: Grundlagen für den professionellen Einsatz*, 176–191. Berlin: BDÜ Fachverlag.
Hasler, Eva, Adria Gispert, Gonzalo Iglesias & Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In Marilyn Walker, Heng Ji & Amanda Stent (Hrsg.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Bd. 2: Short Papers, 506–512. New Orleans, LA: Association for Computational Linguistics. DOI: 10.18653/v1/N18-2081.
Hawkins, John A. 1986. *A comparative typology of English and German: Unifying the contrasts*. London: Croom Helm.
Heine, Antje. 2017. Zwischen Grammatik und Lexikon: Ein forschungsgeschichtlicher Blick auf Funktionsverbgefüge. In *International Conference: Light Verb Constructions in Germanic Languages*. Brussels: Université Saint-Louis. http://centres.fusl.ac.be/LING10_08/document/Linguistics_site/SeSLa/documents/lightverbs/Book%20of%20Abstracts_20171023_FINAL.pdf (11 November, 2018).
Holmback, Heather, Serena Shubert & Jan Spyridakis. 1996. Issues in conducting empirical evaluations of controlled languages. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996)*, 166–177. Leuven: Centre for Computational Linguistics. https://aclanthology.org/www.mt-archive.info/90/CLAW-1996-Holmback.pdf.
Huber, David. 2008. *Precision-based sample size reduction for bayesian experimentation using markov chain simulation*. Los Angeles, CA: University of Southern California. (Diss.).
Huijsen, Willem-Olaf. 1998. Controlled language: An introduction. In *Proceedings of the second international workshop on controlled language application (CLAW 98)*, 1–15. Pittsburg, PA: Language Technologies Institute, Carnegie Mellon University.
Hutchins, John W. 1995. Machine translation: A brief history. In E.F.K. Koerner & R.E. Asher (Hrsg.), *Concise history of the language sciences: from the Sumerians to the cognitivists*, 431–445. Oxford: Pergamon.
Hutchins, John W. 1997. Evaluation of machine translation and translation tools. In Giovanni Battista Varile & Annie Zampolli (Hrsg.), *Survey of the state of the art in human language technology*, Kap. 13.3, 418–419. Pisa: Giardini.

Hutchins, John W. & Herold Somers. 1992. *An introduction to machine translation*. London: Academic Press.

IBM. o.D. *IBM SPSS bootstrapping*. https://www.ibm.com/de-de/marketplace/spss-bootstrapping (12 Juli, 2017).

Isabelle, Pierre, Colin Cherry & George Foster. 2017. A challenge set approach to evaluating machine translation. In Martha Palmer, Rebecca Hwa & Sebastian Riedel (Hrsg.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*, 2486–2496. Copenhagen: Association for Computational Linguistics. DOI: 10.18653/v1/D17-1263.

Isahara, Hitoshi. 1995. JEIDA's test-sets for quality evaluation of MT systems: Technical evaluation from the developer's point of view. In *Proceedings of the Machine Translation Summit V*. Luxembourg. https://aclanthology.org/1995.mtsummit-1.35.

ISO. 2002. *ISO / TR 16982: Ergonomics of human-system interaction: Usability methods supporting human centred design*. Geneva. http://www.iso.org/iso/catalogue_detail?csnumber=31176.

Jacob, Robert J. K. & Keith S. Karn. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In Jukka Hyönä, Ralph Radach & Heiner Deubel (Hrsg.), *The mind's eye: Cognitive and applied aspects of eye movement research*, 573–605. Amsterdam: North Holland. DOI: 10.1016/B978-044451020-4/50031-1.

Jääskeläinen, Riitta. 1993. Investigating translation strategies. In Sonja Tirkkonen-Condit & John Laffling (Hrsg.), *Recent trends in empirical translation research* (Kielitieteellisiä tutkimuksia/Studies in Languages 28), 99–120. Joensuu: University of Joensuu, Faculty of Arts.

Jensen, Kristian Tangsgaard Hvelplund. 2009. Indicators of text complexity. In Susanne Göpferich, Arnt Lykke Jakobsen & Inger M. Mees (Hrsg.), *Behind the mind, methods, models and results in translation process research* (Copenhagen Studies in Language 36), 61–80. Copenhagen: Samfundslitteratur.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. *Google's multilingual neural machine translation system: Enabling zero-shot translation*. DOI: 10.48550/arXiv.1611.04558.

Johnson, R. Burke & Anthony J. Onwuegbuzie. 2004. Mixed methods research: A research paradigm whose time has come. *Educational Researcher* 33(7). 14–26.

Johnson, R. Burke, Anthony J. Onwuegbuzie & Lisa A. Turner. 2007. Towards a definition of mixed methods research. *Journal of Mixed Methods Research* 1(2). 112–133.

Johnston, Bill. 2015. *Survey question guide: Writing scale questions*. https://www.surveygizmo.com/survey-blog/question-scale-length/ (7 Juni, 2016).

Kalchbrenner, Nal & Phil Blunsom. 2013. Recurrent continuous translation models. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu & Steven Bethard (Hrsg.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. Seattle, WA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D13-1176.

Kamprath, Christine, Eric Adolphson, Teruko Mitamura & Eric Nyberg. 1998. Controlled language for multilingual document production: Experience with Cataepillar Technical English. In *The Second International Workshop on Controlled Language Applications (CLAW '98)*.

Keller, Daniela. 2015. *Interpretation von Konfidenzintervallen*. https://statistik-und-beratung.de/2012/10/interpretation-von-konfidenzintervallen/.

Keller, Daniela. 2019. *Was ist Bootstrapping?* https://statistik-und-beratung.de/2019/10/was-ist-bootstrapping/.

King, Margaret. 1993. Evaluation of MT software and methods. In *Proceedings of Translating and the Computer 15*, 111–119. London: Aslib. https://aclanthology.org/1993.tc-1.9.

King, Margaret. 1996. On the notion of validity and the evaluation of MT systems. In Harold Somers (Hrsg.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager*, 189–203. Amsterdam: Benjamins. DOI: 10.1075/btl.18.19kin.

King, Margaret & Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation. In Hans Karlgren (Hrsg.), *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Bd. 2, 211–216. Helsinki. https://aclanthology.org/C90-2037.

King, Margaret, Andrei Popescu-Belis & Eduard Hovy. 2003. FEMTI: Creating and using a framework for MT evaluation. In *Proceedings of Machine Translation Summit IX*, 224–231. New Orleans, LA: AMTA. http://www.amtaweb.org/summit/MTSummit/papers.html (4 Oktober, 2017).

Kirchhoff, Katrin, Daniel Capurro & Anne M. Turner. 2014. A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation* 28(1). 1–17. DOI: 10.1007/s10590-013-9140-x.

Kliegl, Reinhold, Ellen Grabner, Martin Rolfs & Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology* 16(1/2). 262–284.

Klubička, Filip, Antonio Toral & Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics* 108. 121–132.

Koby, Geoffrey S., Paul Fields, Daryl Hague & Alan Melby. 2014. Defining translation quality. *Revista Tradumàtica, Traducció i qualitat* 12. 413–420.

Koehn, Philipp. 2017. *Statistical machine translation, draft of chapter 13: neural machine translation*. DOI: 10.48550/arXiv.1709.07809.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Sophia Ananiadou (Hrsg.), *Proceedings of the ACL-2007 demo and poster sessions*, 177–180. Prague: Association for Computational Linguistics. https://aclanthology.org/P07-2045.

Koehn, Philipp, Franz Josef Och & Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 127–133. https://aclanthology.org/N03-1017.

Koh, Sungryong, Jinee Maeng, Ji-Young Lee, Young-Sook Chae & Key-Sun Choi. 2001. A test suite for evaluation of English-to-Korean machine translation systems. In Bente Maegaard (Hrsg.), *Proceedings of the Machine Translation Summit VIII*. Santiago de Compostela. https://aclanthology.org/2001.mtsummit-papers.35.

Köhn, Philipp. 2010. *Statistical machine translation*. Cambridge: Cambridge University Press.

König, Ekkehard & Volker Gast. 2012. *Understanding English-German contrasts*. Berlin: Erich Schmidt.

Königs, Karin. 2004. *Übersetzen Englisch – Deutsch: Ein systematischer Ansatz*. München: Oldenbourg Wissenschaftsverlag.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia (Hrsg.), *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 181–190. Montreal: Association for Computational Linguistics. https://aclanthology.org/W12-3123.

Krings, Hans. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent, OH: Kent State University Press.

Kuckartz, Udo. 2014. *Mixed Methods, Methodologie, Forschungsdesigns und Analyseverfahren*. Berlin: Springer.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159–174.

Langer, Inghard, Friedemann Schulz von Thun & Reinhard Tausch. 1974. *Verständlichkeit in Schule, Verwaltung, Politik, Wissenschaft*. 10. Aufl. 2015 unter dem Titel: Sich verständlich ausdrücken. München: Reinhardt.

Lavie, Alon. 2010. Evaluating the output of machine translation systems. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Tutorials*. Denver, CO: Association for Machine Translation in the Americas. https://aclanthology.org/2010.amta-tutorials.4.

LDC. 2002. *Linguistic data annotation specification: Assessment of fluency and adequacy in Chinese-English translations*. 1.0. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Le, Quoc & Mike Schuster. 2016. A neural network for machine translation, at production scale. *Google Research Blog*. https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html.

Lehmann, Sabine, Ben Gottesman, Robert Grabowski, Mayo Kudo, Siu Kei Pepe Lo, Melanie Siegel & Frederik Fouvry. 2012. Applying CNL authoring support to improve machine translation of forum data. In Tobias Kuhn & Norbert E. Fuchs (Hrsg.), *Proceedings of the Third International Workshop on Controlled Natural Language (CNL) 2012*, 1–10. Zürich: Springer. DOI: 10.1007/978-3-642-32612-7_1.

Lehrndorfer, Anne. 1996a. Kontrollierte Sprache für Technische Dokumentation: Ein Ansatz für das Deutsche. In Hans P. Krings (Hrsg.), *Wissenschaftliche Grundlagen der technischen Kommunikation*, 339–368. Tübingen: Gunter Narr.

Lehrndorfer, Anne. 1996b. *Kontrolliertes Deutsch: Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der Technischen Dokumentation*. Tübingen: Gunter Narr.

Lehrndorfer, Anne & Ursula Reuther. 2008. *Kontrollierte Sprache: standardisierte Sprache?* Muthig Jürgen (Hrsg.). Bd. 16. Lübeck: Schmidt-Römhild (tekom Hochschulschriften). 97–121.

Lehrndorfer, Anne & Stefanie Schachtl. 1998. Controlled Siemens documentary German and TopTrans. *Technical Communicators, TC-Forum* (3). 8–10.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8). 707–710.

Ley, Martin. 2005. *Kontrollierte Textstrukturen: Ein (linguistisches) Informationsmodell für die technische Kommunikation*. Gießen: Justus-Liebig-Universität Gießen. (Diss.).

Lin, Chin-Yew & Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings NAACL*, 150–157. https://aclanthology.org/N03-1020.

Lin, Chin-Yew & Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 04)*, 605–612. Barcelona: Association for Computational Linguistics. DOI: 10.3115/1218955.1219032.

Liu, Chang, Daniel Dahlmeier & Hwee Tou Ng. 2011. Better evaluation metrics lead to better machine translation. In Regina Barzilay & Mark Johnson (Hrsg.), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 375–384. Edinburgh: Association for Computational Linguistics. https://aclanthology.org/D11-1035.

Liu, Di, Randolph G. Bias, Matthew Lease & Rebecca Kuipers. 2012. Crowdsourcing for usability testing. In *ASIST 2012*. Baltimore, MD. DOI: 10.1002/meet.14504901100.

Llitjós, Ariadna Font, Jaime G. Carbonell & Alon Lavie. 2005. A framework for interactive and automatic refinement to transfer-based machine translation. In Bente Maegaard (Hrsg.), *10th EAMT conference Practical applications of machine translation*, 87–96. Budapest: European Association for Machine Translation. https://aclanthology.org/2005.eamt-1.13.

Lommel, Arle, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis & Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In Mauro Cettolo, Marcello Federico, Lucia Specia & Andy Way (Hrsg.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT-14)*, 165–172. Dubrovnik: European Association for Machine Translation. https://aclanthology.org/2014.eamt-1.38.

Lommel, Arle, Aljoscha Burchardt & Hans Uszkoreit. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of ASLIB Translating and the Computer 34*. https://aclanthology.org/www.mt-archive.info/10/Aslib-2013-Lommel.pdf.

Luong, Minh-Thang, Hieu Pham & Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch & Jian Su (Hrsg.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1412–1421. Lisbon: Association for Computational Linguistics. DOI: 10.18653/v1/D15-1166.

Macketanz, Vivien, Renlong Ai, Aljoscha Burchardt & Hans Uszkoreit. 2018. TQ-AutoTest: An automated test suite for (machine) translation quality. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga

(Hrsg.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 886–892. Miyazaki: European Language Resources Association (ELRA). https://aclanthology.org/L18-1142.

Martin, Juan Alberto Alonso & Anna Civil Serra. 2014. Integration of a machine translation system into the editorial process flow of a daily newspaper. *Procesamiento del Lenguaje Natural* 53. 193–196.

Matusov, Evgeny. 2019. The challenges of using neural machine translation for literature. In James Hadley, Maja Popović, Haithem Afli & Andy Way (Hrsg.), *Proceedings of the qualities of literary machine translation*, 10–19. Dublin: European Association for Machine Translation. https://aclanthology.org/W19-7302.

McCord, Michael C. 1989. Design of LMT: A prolog-based machine tranalation system. *Computational Linguistics* 15(1). 33–52.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

McMurrey, David A. 2006. *Online technical writing course guide, section: highlighting & emphasis*. https://www.tu-chemnitz.de/phil/english/sections/linguist/independent/kursmaterialien/TechComm/acchtml/acctoc.html (10 November, 2017).

Mehta, Sneha, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani & Ritwik Kumar. 2020. Simplify-then-translate: Automatic preprocessing for black-box translation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York: AAAI Press. DOI: 10.1609/aaai.v34i05.6369.

Mertin, Elvira. 2006. *Prozessorientiertes Qualitätsmanagement im Dienstleistungsbereich Übersetzen* (Leipziger Studien zur angewandten Linguistik und Translatologie 2). Zugl.: Universität Leipzig, Dissertation. Frankfurt a. M.: Lang.

Mitamura, Teruko. 1999. Controlled language for multilingual machine translation. In *Proceedings of Machine Translation Summit VII: MT in the Great Translation Era*, 46–52. Singapore. https://aclanthology.org/1999.mtsummit-1.8.

Mitchell, Linda, O'Brien Sharon & Johann Roturier. 2014. Quality evaluation in community post-editing. *Machine Translation* 28(3). 237–262. DOI: 10.1007/s10590-014-9160-1.

Møller, Margrethe H. 2003. Grammatical metaphor, controlled language and machine translation. In *Proceedings of EAMT-CLAW-03*, 95–104. Dublin: Dublin City University. https://aclanthology.org/2003.eamt-1.11.

Molnár, Ondře. 2012. Source text quality in the translation process. In Jitka Zehnalová, Ondře Molnár & Michal Kubánek (Hrsg.), *Tradition and trends in trans-language communication*, 59–86. Olomouc: Palacký University.

Mügge, Uwe. 2013. Implementing a controlled language is now cheaper and easier than ever. *tcworld*. http://works.bepress.com/uwe_muegge/91/ (25 März, 2015).

Müller, Mathias, Annette Rios, Elena Voita & Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi & Karin Verspoor (Hrsg.), *Proceedings of WMT 2018*, 61–72. Brussels: Association for Computational Linguistics. DOI: 10.18653/v1/W18-6307.

Muthig, Jürgen. 2003. Das Passiv: Über den Umgang mit dem Passiv. In *Technische Dokumentation 2003/07_08*. http://www.doku.net/artikel/daspassiv.htm (13 November, 2017).

Ñeco, Ramón P. & Mikel L. Forcada. 1997. Asynchronous translations with recurrent neural nets. *Neural Networks* 4. 2535–2540.

Nerius, Dieter. 2007. *Deutsche Orthographie*. 4. Aufl. Hildesheim: Olms.

Ng, Vincent. 2017. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4877–4884. San Francisco, CA: AAAI Press. DOI: 10.1609/aaai.v31i1.11149.

Nießen, Sonja, Franz Josef Och, Gregor Leusch & Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (Hrsg.), *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens: European Language Resources Association (ELRA). https://aclanthology.org/L00-1210/.

Nitzke, Jean. 2019. *Problem solving activities in post-editing and translation from scratch: A multi-method study*. Berlin: Language Science Press. DOI: 10.5281/zenodo.2546446.

Nyberg, Eric & Teruko Mitamura. 1996. Controlled language and knowledge-based machine translation: Principles and practice. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996)*, 74–83. Leuven: Centre for Computational Linguistics. https://aclanthology.org/www.mt-archive.info/90/CLAW-1996-Nyberg.pdf.

Nyberg, Eric, Teruko Mitamura & Willem-Olaf huijsen. 2003. Controlled language for authoring and translation. In Harold Somers (Hrsg.), *Computers and translation: A translator's guide*, 245–281. Amsterdam: Benjamins.

O'Brien, Sharon. 2003. Controlling controlled English: An analysis of several controlled English rule sets. In *Proceedings of the 4th Controlled Language Applications Workshop (CLAW)*, 105–114. Dublin.

O'Brien, Sharon. 2006. *Machine translatability and post-editing effort: An empirical study using translog and choice network analysis*. Dublin: Dublin City University. (Diss.).

O'Brien, Sharon. 2010a. Controlled language and readability. In Gregory M. Shreve & Erik Angelone (Hrsg.), *Translation and cognition* (American Translators Association Scholarly Monograph 15), 143–165. Amsterdam: Benjamins. DOI: 10.1075/ata.xv.08obr.

O'Brien, Sharon. 2010b. Introduction to post-editing: Who, what, how and where to next. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Tutorials*. Denver, CO: Association for Machine Translation in the Americas. https://aclanthology.org/2010.amta-tutorials.1.

O'Brien, Sharon. 2011. Towards predicting post-editing productivity. *Machine Translation* 3. 197–215.

O'Brien, Sharon & Johann Roturier. 2007. How portable are controlled languages rules: a comparison of two empirical MT studies. In Bente Maegaard (Hrsg.), *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*. Copenhagen. https : / / aclanthology.org/2007.mtsummit-papers.46.

O'Brien, Sharon, Johann Roturier & Giselle de Almeida. 2009. Post-editing MT output: Views from the researcher, trainer, publisher and practitioner. In *Machine Translation Summit XII*. Ottawa.

Och, Franz Josef & Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics* 30(4). 417–449. DOI: 10.1162/0891201042544884.

Ogden, Charles Kay. 1935. *The A B C of basic English (in basic)*. London.

Ogden, Charles Kay & Ivor Armstrong Richards. 1923. *The meaning of meaning*. New York: Harvest/HBJ.

Padó, Sebastian, Daniel M. Cer, Michel Galley, Dan Jurafsky & Christopher D. Manning. 2009. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation* 23(2–3). 181–193.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, PA: Association for Computational Linguistics. DOI: 10 . 3115 / 1073083.1073135.

Petrzelka, Jiri. 2011. *Statistical machine translation: How to maximize the quality of translation*. Saarbrücken: LAP Lambert Academic Publishing.

Pickering, Martin J. & Matthew J. Traxler. 1998. Plausibility and recovery from garden paths: An eye tracking study. *Journal of Experimental Psychology Learning, Memory, and Cognition* 24. 940–961.

Poibeau, Thierry. 2017. *Machine translation*. Cambridge, MA: MIT Press.

Popović, Maja. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics* 96. 59–68.

Popović, Maja. 2017. Comparing language related issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics* 108(1). 209–220.

Popović, Maja. 2018. Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation* 32(3). 237–253. DOI: 10.1007/s10590-018-9219-5.

Popović, Maja, Mihael Arcan & Arle Lommel. 2016. Potential and limits of using post-edits as reference translations for MT evaluation. *Baltic Joural of Modern Computing* 4(2). 218–229.

Popović, Maja & Aljoscha Burchardt. 2011. From human to automatic error classification for machine translation output. In Mikel L. Forcada, Heidi Depraetere & Vincent Vandeghinste (Hrsg.), *Proceedings of EAMT*, 265–272. Leuven: European Association for Machine Translation. https://aclanthology.org/2011.eamt-1.36.

Popović, Maja, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis & Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In Marko Tadic, Philipp Koehn, Johann Roturier & Andy Way (Hrsg.), *Proceedings of the 17th annual conference of the European association for machine translation (EAMT 14)*, 191–198. Dubrovnik: European Association for Machine Translation. https://aclanthology.org/2014.eamt-1.41.

Popović, Maja & Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics* 37(4). 657–688.

Porst, Rolf. 2011. *Fragebogen: Ein Arbeitsbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften. DOI: 10.1007/978-3-531-92884-5.

Potet, Marion, Emmanuelle Esperança-Rodier, Laurent Besacier & Hervé Blanchon. 2012. Collection of a large database of French-English SMT output corrections. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Eighth International Conference*

on Language Resources and Evaluation (LREC'12), 4043–4048. Istanbul: ELRA. http://www.lrec-conf.org/proceedings/lrec2012/pdf/506_Paper.pdf.

Pouget-Abadie, Jean, Dzmitry Bahdanau, Bart van Merrienboer, Kyunghyun Cho & Yoshua Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In Dekai Wu, Marine Carpuat, Xavier Carreras & Eva Maria Vecchi (Hrsg.), *Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 78–85. Doha: Association for Computational Linguistics. DOI: 10.3115/v1/W14-4009.

Püschel, Ulrich. 1996. Sprachstil: ein Thema für Technische Redakteure? In Hans P. Krings (Hrsg.), *Wissenschaftliche Grundlagen der technischen Kommunikation*, 339–368. Tübingen: G. Narr.

Ramírez Polo, Laura. 2012. *Use and evaluation of controlled languages in industrial environments and feasibility study for the implementation of machine translation*. Valencia: Universidad de Valencia. (Diss.).

Ramírez Polo, Laura & Johann Haller. 2005. Controlled language and the implementation of machine translation for technical documentation. In *Proceedings of the 27th International Conference on Translating and the Computer*. London: Aslib. https://aclanthology.org/2005.tc-1.13.

Ramlow, Markus. 2008. Maschinelle Übersetzungssysteme im Vergleich. In Michael Krenz, Markus Ramlow & Uta Seewald-Heeg (Hrsg.), *Maschinelle Übersetzung und XML im Übersetzungsprozess: Prozesse der Translation und Lokalisierung im Wandel*, 15–150. Berlin: Frank & Timme.

Rascu, Ecaterina. 2006. A controlled language approach to text optimisation in technical documentation. In Miriam Butt (Hrsg.), *Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Sprache)*, 107–114. Konstanz: Universität Konstanz.

Rehbein, Jochen. 1988. Ausgewählte Aspekte der Pragmatik. In Ulrich Ammon, Norbert Dittmar & Klaus Mattheier (Hrsg.), *Soziolinguistik*, 1181–1195. Berlin: De Gruyter.

Reuther, Ursula. 2003. Two in one: Can it work? Readability and translatability by means of controlled language. In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop (CLAW 2003)*, 124–132. Budapest: European Association for Machine Translation. https://aclanthology.org/2003.eamt-1.14.

Reuther, Ursula. 2007. Controlled language! Controlled translation? Präsentation auf dem CAT Workshop "The Next Chapter", Mai 2007, Jülich.

Rikters, Matiss & Ondrej Bojar. 2019. *Paying attention to multi-word expressions in neural machine translation*. DOI: 10.48550/arXiv.1710.06313.

Rösener, Christoph. 2010. Computational linguistics in the translator's workflow: Combining authoring tools and translation memory systems. In Michael Piotrowski, Cerstin Mahlow & Robert Dale (Hrsg.), *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing*, 1–6. Los Angeles, CA: Association for Computational Linguistics. https://aclanthology.org/W10-0401.

Roturier, Johann. 2006. *An investigation into the impact of controlled English rules on the comprehensibility, usefulness, and acceptability of machine-translated technical documentation for French and German users*. Dublin: Dublin City University. (Diss.).

Roturier, Johann, Linda Mitchell, Robert Grabowski & Melanie Siegel. 2012. Using automatic machine translation metrics to analyze the impact of source reformulations. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas (AMTA): Research Papers*. San Diego, CA: Association for Machine Translation in the Americas. https://aclanthology.org/2012.amta-papers.24/.

Sager, Juan C. 1994. *Language engineering and translation: Consequences of automation* (Benjamins Translation Library 1). Amsterdam: Benjamins.

Saldanha, Gabriela & Sharon O'Brien. 2014. *Research methodologies in translation studies*. New York: Routledge.

Schenk, Eric & Claude Guittard. 2011. Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics & Management* 7(1). 93–107. DOI: 10.3917/jie.007.0093.

Schottmüller, Nina & Joakim Nivre. 2014. Issues in translating verb-particle constructions from German to English. In Valia Kordoni, Markus Egg, Agata Savary, Eric Wehrli & Stefan Evert (Hrsg.), *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 124–131. Gothenburg: Association for Computational Linguistics. DOI: 10.3115/v1/W14-0821.

Schubert, Klaus. 1999. Zur Automatisierbarkeit des Übersetzens. In Alberto Gil, Johann Haller, Erich Steiner & Heidrun Gerzymisch-Arbogast (Hrsg.), *Modelle der Translation* (Sabest Saarbrücker Beiträge zur Sprach- und Translationswissenschaft 1), 423–441. Frankfurt am Main: Lang.

Schwanke, Martina. 1991. *Maschinelle Übersetzung: ein Überblick über Theorie und Praxis*. Berlin: Springer.

Schwitter, Rolf. 2007. *Controlled natural languages*. Techn. Ber. Centre for Language Technology, Macquarie University.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd & Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32(3). 217–235.

Shubert, Serena K., Jan H. Spyridakis, Heather K. Holmback & Mary B. Coney. 1995. The comprehensibility of simplified English in procedures. In *Proceedings of the Professional Communication Conference: 'Smooth sailing to the Future', IEEE International*, 171–173.

Siegel, Melanie. 2011. Autorenunterstützung für die Maschinelle Übersetzung. In Hanna Hedeland, Thomas Schmidt & Kai Wörne (Hrsg.), *Multilingual resources and multilingual application: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, 183–186. Hamburg: Unviersität Hamburg. http://exmaralda.org/gscl2011/downloads/AZM96.pdf.

Siegel, Melanie. 2013. Authoring support for controlled language and machine translation. In Georg Rehm, Felix Sasaki, Daniel Stein & Andreas Witt (Hrsg.), *Translation: Computation, corpora, cognition. Special issue on language technologies for a multilingual Europe*, Bd. 3, 49–60. https://www.blogs.uni-mainz.de/fb06-tc3/files/2015/11/29-146-1-PB.pdf.

Siegel, Melanie. 2014. Englisch nach Regelwerk. *Technische Kommunikation* (4). 28–32.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, 223–231. Cambridge, MA: Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky & Andrew Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In Mirella Lapata & Hwee Tou Ng (Hrsg.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, 254–263. Honolulu, HI: Association for Computational Linguistics. https://aclanthology.org/D08-1027.

Specia, Lucia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In Mikel L. Forcada, Heidi Depraetere & Vincent Vandeghinste (Hrsg.), *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, 73–80. Leuven: European Association for Machine Translation. https://aclanthology.org/2011.eamt-1.12.

Specia, Lucia, Nicola Cancedda & Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'2010)*, 3375–3378. Valletta: Euro-

pean Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/504_Paper.pdf.

Spyridakis, Jan, Serena Shubert & Heather Holmback. 1997. Measuring the translatability of simplified English in procedural documents. *IEEE Transactions on Professional Communication* 40(1).

Spyridakis, Jan, Carolyn Wei, Jennifer Barrick, Elisabeth Cuddihy & Brandon Maust. 2005. Internet-based research: Providing a foundation for web-design guidelines. *IEEE Transactions on Professional Communication* 48(3). 242–260.

Statista. 2019. *Rangfolge der wichtigsten Handelspartner Deutschlands nach Wert der Exporte im Jahr 2018*. https://de.statista.com/statistik/daten/studie/2876/umfrage/rangfolge-der-wichtigsten-handelspartner-deutschlands-nach-wert-der-exporte/ (11 November, 2019).

Stein, Daniel. 2009. Maschinelle Übersetzung: ein Überblick. *Journal for Language Technology and Computational Linguistics* 3. 5–18.

Stevens, Darren. 2018. *Crowdsourcing: Pros, cons, and more*. https://www.hongkiat.com/blog/what-is-crowdsourcing/ (26 Februar, 2019).

Stoessel, Sabine. 2002. Methoden des Testings im Usability Engineering. In Markus Beier & Vittoria von Gizycki (Hrsg.), *Usability: Nutzerfreundliches Web-Design*, 75–96. Berlin: Springer.

Stojanovski, Dario & Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi & Karin Verspoor (Hrsg.), *Proceedings of the Third Conference on Machine Translation (WMT)*, Bd. 1: Research Papers, 49–60. Brussels: Association for Computational Linguistics. DOI: 10.18653/v1/W18-6306.

Stojanovski, Dario & Alexander Fraser. 2019. Improving anaphora resolution in neural machine translation using curriculum learning. In Mikel Forcada, Andy Way, Barry Haddow & Rico Sennrich (Hrsg.), *Proceedings of the Machine Translation Summit XVII: Research Track*, 140–150. Dublin: European Association for Machine Translation. https://aclanthology.org/W19-6614.

Stymne, Sara. 2011. Blast: A tool for error analysis of machine translation output. In Sadao Kurohashi (Hrsg.), *Proceedings of the 49th ACL*, 56–61. Portland, OR: Association for Computational Linguistics. https://aclanthology.org/P11-4010.

Stymne, Sara. 2013. Using a grammar checker and its error typology for annotation of statistical machine translation errors. In Jani-Matti Tirkkonen & Esa Anttikoski (Hrsg.), *Proceedings of the 24th Scandinavian Conference of Linguistics*, 332–344. Joensuu: University of Eastern Finland.

Su, Keh-Yih, Ming-Wen Wu & Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation system. In *COLING '92: Proceedings of the 14th conference on Computational linguistics*, Bd. 2, 433–439. Nantes: Association for Computational Linguistics. DOI: 10.3115/992133.992137.

Sutskever, Ilya, Oriol Vinyals & Quoc V. Le. 2014. *Sequence to sequence learning with neural networks*. DOI: 10.48550/arXiv.1409.3215.

Swan, Michael. 1980. *Practical English usage*. Oxford: Oxford University Press.

Tatsumi, Midori. 2009. Correlation between automatic evaluation metric scores, post-editing speed and some other factors. In *Proceedings of Machine Translation Summit XII: Posters*, 332–339. Ottawa. https://aclanthology.org/2009.mtsummit-posters.20.

Tatsumi, Midori & Johann Roturier. 2010. *Source text characteristics and technical and temporal post-editing effort: what is their relationship?* Ventsislav Zhechev (Hrsg.). Denver, CO: Association for Machine Translation in the Americas. 43–51. https://aclanthology.org/2010.jec-1.6.

TAUS Report. 2010. *Postediting in practice*. https://www.taus.net/reports/postediting-in-practice.

Teich, Elke. 2003. *Cross-linguistic variation in system and text*. Berlin: Mouton de Gruyter.

tekom RG Alb Donau. 2010. *Maschinenrichtlinie 2006/42/EG: Herausforderung Dokumentations-Bevollmächtigter/Bevollmächtigte: Was kommt auf diese Person zu?* https://webforum.tekom.de/fileadmin/user_upload/tekom/berichte/uploaded_file1024.pdf (11 November, 2019).

Temnikova, Irina. 2010. Cognitive evaluation approach for a controlled language post-editing experiment. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 3485–3490. Valetta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/437_Paper.pdf.

Thicke, Lori. 2011. Improving MT results: A study. *Multilingual* 22(1). 37–40.

Tomita, Masaru, Masako Shirai, Junya Tsutsumi, Miki Matsumura & Yuki Yoshikawa. 1993. Evaluation of MT systems by TOEFL. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, 252–265. Kyoto. https://aclanthology.org/1993.tmi-1.22.

Toral, Antonio & Victor M. Sanchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In Mirella Lapata, Phil Blunsom & Alexander Koller (Hrsg.), *Proceedings of the 15th conference of the European chapter of the association for computational*

linguistics, Bd. 1: Long Papers, 1063–1073. Valencia: Association for Computational Linguistics. https://aclanthology.org/E17-1100.

Toral, Antonio & Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In Joss Moorkens, Sheila Castilho, Federico Gaspari & Stephen Doherty (Hrsg.), *Translation quality assessment: From principles to practice*, 263–287. Cham: Springer.

Toral, Antonio, Martijn Wieling & Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities* 5(9). DOI: 10.3389/fdigh.2018.00009.

Tripathi, Sneha & Juran Krishna Sarkhel. 2010. Approaches to machine translation. *Annals of Library and Information Studies* 57. 388–393.

Underwood, Nancy & Bart Jongejan. 2001. Translatability checker: A tool to help decide whether to use MT. In Bente Maegaard (Hrsg.), *Proceedings of the Machine Translation summit VII: Machine Translation in the Information Age*, 363–368. Santiago de Compostela: Center for Sprogteknologi. https://aclanthology.org/2001.mtsummit-papers.65.

Van Brussel, Laura, Arda Tezcan & Lieve Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (Hrsg.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1600.

Van der Eijk, Pim, Michiel de Koning & Gert van der Steen. 1996. Controlled language correction and translation. In *Proceedings of the First International Workshop on Controlled Language Applications (CLAW)*, 64–73. Leuven. https://aclanthology.org/www.mt-archive.info/90/CLAW-1996-vanderEijk.pdf.

Van Gog, Tamara, Kester Liesbeth, Fleurie Nievelstein, Bas Giesbers & Fred Paas. 2009. Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior* 25(2). 325–331.

Van Slype, Georges. 1979. *Critical study of methods for evaluating the quality of machine translation.* Techn. Ber. BR 19142. Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Luxembourg.

Vanni, Michelle & Keith Miller. 2002. Scaling the ISLE framework: Use of existing corpus resources for validation of MT evaluation metrics across languages. In

Proceedings of LREC 2002. Las Plamas: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/306.pdf.

Vardaro, Jennife, Moritz Schaeffer & Silvia Hansen-Schirra. 2019. Translation quality and error recognition in professional neural machine translation post-editing. *Informatics* 6(Article 41). DOI: 10.3390/informatics6030041.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. *Attention is all you need*. DOI: 10.48550/arXiv.1706.03762.

Vauquois, Bernard. 1968. A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress 68*, Bd. 2, 254–260. Edinburgh.

Vilar, David, Jia Xu, Luis Fernando D'Haro & Hermann Ney. 2006. Error analysis of machine translation output. In *LREC-2006: Proceesings of the 5th international conference on language resources and evaluation*, 697–702. Genoa: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf.

Voita, Elena, Rico Sennrich & Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In Kentaro Inui, Jing Jiang, Vincent Ng & Xiaojun Wan (Hrsg.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 877–886. Hong Kong: Association for Computational Linguistics. DOI: 10.18653/v1/D19-1081.

Voita, Elena, Pavel Serdyukov, Rico Sennrich & Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, 1264–1274. Melbourne: Association for Computational Linguistics. DOI: 10.18653/v1/P18-1117.

Volk, Martin. 2018. Parallel corpora, terminology extraction and machine translation. In *16. DTT-Symposion, Terminologie und Text(e)*, 3–14. Mannheim. DOI: 10.5167/uzh-150769.

Wagner, Emma. 1987. Post-editing: Practical considerations. In Catriona Picken (Hrsg.), *The business of translation and interpreting*, 71–78. London: Aslib.

Way, Andrew. 2010. Machine translation. In Alexander Clark, Chris Fox & Shalom Lappin (Hrsg.), *The handbook of computational linguistics and natural language processing*, 531–573. Malden: Wiley-Blackwell. DOI: 10.1002/9781444324044.ch19.

Weber, Nico (Hrsg.). 1998. *Machine translation: Theory, applications, and evaluation* (Sprachwissenschaft, Computerlinguistik und neue Medien 1). St. Augustin: Gardez!-Verlag.

Wells Akis, Jennifer, Stephanie Brucker, Virginia Chapman, Layne Ethington, Bob Kuhns & PJ Schemenaur. 2003. Authoring translation-ready documents: Is software the answer? Can controlled languages scale to the web? In *Proceedings of the 21st Annual International Conference on Documentation (SIGDOC 2003)*, 39–44. San Francisco, CA: Association for Computing Machinery. DOI: 10.1145/944868.944878.

Werthmann, Antonina & Andreas Witt. 2014. Maschinelle Übersetzung: Gegenwart und Perspektiven. In Gerhard Stickel (Hrsg.), *Translation and interpretation in Europe: Contributions to the Annual Conference 2013 of EFNIL in Vilnius*, 79–103. Frankfurt am Main: Lang.

White, John. 2003. How to evaluate machine translation. In Harold Somers (Hrsg.), *Computers and translation: A translator's guide*, 211–244. Amsterdam: Benjamins.

White, John S. & Theresa A. O'Connell. 1996. Adaptation of the DARPA machine translation evaluation paradigm to end-to-end systems. In *Expanding MT Horizons, Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, 106–114. Montreal. http://aclanthology.lst.uni-saarland.de/1996.amta-1.11.pdf.

Wisniewski, Guillaume, Anil Kumar Singh, Natalia Segal & François Yvon. 2013. Design and analysis of a large corpus of post-edited translations: Quality estimation, failure analysis and the variability of post-edition. In Andy Way, Khalil Sima'an & Mikel L. Forcada (Hrsg.), *Proceedings of Machine Translation Summit XIV*, 117–124. Nice. https://aclanthology.org/2013.mtsummit-papers.15.

Wittkowsky, Marion. 2017. Regulierte Sprache und (maschinelle) Übersetzung in der Fachkommunikation. In Jörg Porsiel (Hrsg.), *Maschinelle Übersetzung: Grundlagen für den professionellen Einsatz*, 84–94. Berlin: BDÜ-Fachverlag.

Wong, Billy & Chun yu Kit. 2009. Atec: Automatic evaluation of machine translation via word choice and word order. *Machine Translation* 23(2–3). 141–155.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. *Google's neural machine translation system: bridging the gap between human and machine translation.* DOI: 10.48550/arXiv.1609.08144.

Xing, Shi, Kevin Knight & Deniz Yuret. 2016. Why neural translations are the right length. In Jian Su, Kevin Duh & Xavier Carreras (Hrsg.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2278–

2282. Austin, TX: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1248.
Zaninello, Andrea & Alexandra Birch. 2020. Multiword expression aware neural machine translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 3816–3825. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.471.
Zeman, Daniel, Mark Fishelb, Jan Berkaa & Ondřej Bojara. 2011. Addicter: What is wrong with my translations? *Bulletin of Mathematical Linguistics* 96. 79–88.
Zhang, Jiajun & Chengqing Zong. 2020. *Neural machine translation: Challenges, progress and future*. DOI: 10.48550/arXiv.2004.05809.

Autorenregister

Sprachkontrolle im Spiegel der Maschinellen Übersetzung

Examining the general impact of the Controlled Languages rules in the context of Machine Translation has been an area of research for many years. The present study focuses on the following question: How do the Controlled Language (CL) rules impact the Machine Translation (MT) output individually? Analyzing a German corpus-based test suite of technical texts that have been translated into English by different MT systems, the study endeavors to answer this question at different levels: the general impact of CL rules (rule- and system-independent), their impact at rule level (system-independent), their impact at system level (rule-independent), and at rule and system level. The results of five MT systems (a rule-based system, a statistical system, two differently constructed hybrid systems, and a neural system) are analyzed and contrasted. For this, a mixed-methods triangulation approach that includes error annotation, human evaluation, and automatic evaluation was applied. The data were analyzed both qualitatively and quantitatively based on the following parameters: number and type of MT errors, style and content quality, and scores from two automatic evaluation metrics. In line with many studies, the results show a general positive impact of the applied CL rules on the MT output. However, at rule level, only four rules proved to have positive effects on all parameters; three rules had negative effects on the parameters; and two rules did not show any significant impact. At rule and system level, the rules affected the MT systems differently, as expected. Some rules that had a positive impact on earlier MT approaches did not show the same impact on the neural MT approach. Furthermore, the neural MT delivered distinctly better results than earlier MT approaches, namely the highest error-free, style and content quality rates both before and after the rules application, which indicates that the neural MT offers a promising solution that no longer requires CL rules for improving the MT output, what in turn allows for a more natural style.