COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 1, 19–38
https://doi.org/10.1080/19312458.2021.1955845

Routledge
Taylor & Francis Group

OPEN ACCESS

Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections

Daniel Maier[a], Christian Baden[b], Daniela Stoltenberg[a], Maya De Vries-Kedem[c], and Annie Waldherr[d]

[a]Institute for Media and Communication Studies, Free University of Berlin, Berlin, Germany; [b]Department of Communication and Journalism, Hebrew University of Jerusalem, Jerusalem, Israel; [c]Department of Communication and Journalism & Swiss Center for Conflict Research, Management and Resolution, Hebrew University of Jerusalem, Jerusalem, Israel; [d]Department of Communication, University of Vienna, Wien, Austria

ABSTRACT
The goal of this paper is to evaluate two methods for the topic modeling of multilingual document collections: (1) machine translation (MT), and (2) the coding of semantic concepts using a multilingual dictionary (MD) prior to topic modeling. We empirically assess the consequences of these approaches based on both a quantitative comparison of models and a qualitative validation of each method's potentials and weaknesses. Our case study uses two text collections (of tweets and news articles) in three languages (English, Hebrew, Arabic), covering the ongoing local conflicts between Israeli authorities, settlers, and Palestinian Bedouins in the West Bank. We find that both methods produce a large share of equivalent topics, especially in the context of fairly homogenous news discourse, yet show limited but systematic differences when applied to highly heterogenous social media discourse. While the MD model delivers a more nuanced picture of conflict-related topics, it misses several more peripheral topics, especially those unrelated to the dictionary's focus, which are picked up by the MT model. Our study is a first step toward instrument validation, indicating that both methods yield valid, comparable results, while method-specific differences remain.

The necessity for analytic techniques suitable to explore large-scale multilingual text collections has never been more pressing in communication research. Not only are we confronted with an exponential growth of digital communication traces on the Web and social networking sites all over the world, speaking in a multiplicity of languages (e.g., Lazer et al., 2009); but following the rise of transnational networks and globalized cross-border communication (Castells, 2008; Volkmer, 2014), communication on digital media is increasingly diluting linguistic barriers, as the same issues are debated in many locales and multiple languages at once (Baden et al., 2021; Lück et al., 2018).

To date, however, most available methods for computational text analysis are developed to operate on monolingual corpora (Reber, 2019). This is particularly true for topic modeling (Blei et al., 2003), a method for computational text analysis that has become increasingly popular among communication scholars. Topic models are statistical models that are used to extract thematically coherent word clusters, called topics, from document collections. These models build on the assumption that semantically similar words tend to be used in similar contexts (Boyd-Graber & Blei, 2009) – an assumption that fails for multilingual text collections. Since semantically similar words in different languages will occur in largely mutually exclusive contexts, resulting topic models will separate the

CONTACT Christian Baden c.baden@mail.huji.ac.il Department of Communication and Journalism, Mount Scopus, Hebrew University of Jerusalem, Jerusalem 91905, Israel

Supplemental data for this article can be accessed on the publisher's website.

topic space along languages contained in the corpus (Reber, 2019). Metaphorically speaking, the vocabularies of different languages lack a "common denominator."

To date, there are three dominant strategies (Lind et al., 2019a) to avoid this undesired "confusion of tongues" (Reber, 2019, p. 103): (1) calculating separate, language-specific models, one for each language compartment of the corpus (e.g., Heidenreich et al., 2019); (2) enabling topic models to process multilingual corpora by adding information about cross-lingual equivalencies (Boyd-Graber & Blei, 2009; Chan et al., 2020; Mimno et al., 2009); and (3) translating the documents of a multilingual corpus into one common language prior to topic modeling (for an overview of available methods see Lind et al., 2019a). Among these strategies, the first is unsuitable for the integrated analysis of multilingual discourses and faces severe limitations for cross-lingual comparative analysis, as it cannot distinguish between valid differences and differences introduced by the separately estimated models. By contrast, the latter two strategies bear promise for further development.

In our paper, we ask what difference the choice of strategy makes for the topic modeling of multilingual corpora. However, instead of investigating which strategy performs "better,"[1] we seek to understand *in what ways* topic models differ from one another depending on the strategy chosen to bridge the linguistic gap. Through a systematic, quantitative and qualitative comparison, we evaluate how different strategies yield comparable or different findings, and what advantages and limitations accompany them (Grimmer & Stewart, 2013).

To do so, we compare two maximally different approaches to the topic modeling of multilingual corpora: The first approach, which is located squarely within the third tradition, leverages machine translation (MT) services to automatically transfer documents into a common language (here: English) prior to modeling (see also Lucas et al., 2015; Reber, 2019; De Vries et al., 2018). The second approach is located between the second and third strand: Prior to modeling, it transfers the document content from the (language-dependent) lexical level to a language-independent semantic level, using a multilingual dictionary (MD; e.g., Baden & Stalpouskaya, 2015; Lind et al., 2019b). While the MT approach has already been described and applied (e.g., Lucas et al., 2015; De Vries et al., 2018), combining multilingual dictionaries with topic modeling has not yet been discussed in the literature and is an original contribution of this article.

To examine the comparative strengths and limitations of these two approaches, we use a case study that exemplifies important, yet typical challenges in comparative communication research. Focusing on the ongoing local conflicts between Israeli authorities, settlers, and Palestinian Bedouins in the Israeli-controlled part of the West Bank ("Area C"), we collated two multilingual corpora of news and Twitter discourse, respectively, including content in Hebrew, Arabic, and English language. While most existing research using topic models focuses on English language contents, both Hebrew and Arabic are morphologically rich, Semitic languages that raise numerous additional challenges (Tsarfaty et al., 2013) that have received little attention in the available scholarship (for a notable exception, see Lucas et al., 2015). For each corpus, we estimate a range of topic models, using either MT or MD to bridge the three languages. Subsequently, we combine quantitative and qualitative strategies to identify important similarities and differences between the obtained models and derive recommendations for their use in future research.

The remainder of our manuscript proceeds as follows: We first outline why multilingual text collections are ill-suited for standard procedures and review what approaches exist for overcoming such linguistic barriers. Next, we introduce our case study, as well as the specific research design and methods used for the present effort at comparative validation. Following our presentation of findings, we finally discuss their implications for multilingual topic modeling and computational text analysis.

Topic Models, the Babel Problem and Solution Approaches

We use the term *topic model* for a class of computational content analysis techniques that can be used to identify and describe latent thematic structures, called topics, in large text collections (Blei, 2012). In their seminal paper, Blei et al. (2003) introduced a Bayesian statistical generative process that mimics how documents are created. To model this process, topic models introduce topics as unobserved variables that organize patterns of word occurrence into discrete, weighted word collections so as to represent larger thematic structures in a corpus (Griffiths & Steyvers, 2004). Based on the assumption that semantically similar words tend to be used in semantically similar contexts (Boyd-Graber & Blei, 2009), standard topic modeling algorithms rely on the observed co-occurrence patterns to (re-) construct clusters of semantically related words. Technically speaking, topics are conceptualized as conditional probability distributions over a fixed vocabulary (ϕ) and documents are defined as probability distributions over topics (θ). Due to the statistical setup of the model (i.e., the Dirichlet-prior for θ and ϕ), only a few topics are highly prevalent in most single documents, and only few (about 20–30) words have a high probability for most topics.

In the case of multilingual text collections, standard topic modeling algorithms run into what Chan et al. (2020) refer to as the "Babel problem." Whenever a corpus contains documents written in two or more languages, each language uses a different vocabulary. Although there may be numerous words with equivalent meaning across languages, therefore, these words will appear in completely different vocabulary contexts. For example, the English word "university" and its German equivalent "Universität" can be considered semantically identical, but they will appear in the context of diverging (language-specific) vocabularies that overlap only accidentally when multiple languages use the same words to express the same meaning (i.e., {"library," "book," "graduation," "learn," "*professor*"} vs. {"bibliothek," "buch," "abschluss," "lernen," "*professor*"}).[2] Accordingly, the constitutive assumption that semantically similar words tend to be used in similar contexts holds only for semantically similar words *in the same language*. Due to the extremely limited overlap in the vocabularies of different (even related) languages, topic models are unable to recognize semantically similar words in different languages and will separate the topic space along languages contained in the corpus (Reber, 2019).

This inability of topic models to bridge linguistic gaps severely limits their applicability for the expanding field of comparative communication research (e.g., Heidenreich et al., 2019; Vliegenthart & Damstra, 2018; Baden et al., 2020). The problem especially haunts the use of topic models in the study of social media discourse, which frequently includes contributions in multiple languages united around a common focus and shared hashtags (e.g., Kligler-Vilenchik et al., 2020; Theocharis et al., 2016). As a consequence, the Babel problem is receiving increasing attention within and beyond the social sciences (Lucas et al., 2015). We devote the next four subsections to expanding Lind et al.'s (2019a) typology of approaches to bridging between languages (a) post estimation; (b) by reference to external equivalence information; or (c) using machine translation by adding (d) multilingual dictionaries as a novel approach; additionally, we discuss the main advantages and limitations of each solution.

Bridging Babel Post Estimation

While in many cases calculating separate models is a legitimate procedure, it is accompanied by many methodological problems that limit the comparability of the models. For instance, different languages may possess different structural properties (e.g., morphology, syntax; Tsarfaty et al., 2013), requiring different preprocessing steps to harmonize the data and ensure that detected differences are due to semantic and not syntactic or morphological differences. Even where similar preprocessing steps are suitable, it may be hard to find equivalent high-quality resources (e.g., language-specific parsers or lemmatizers) for each language (Baden et al., 2020). Thematic diversity may differ between included languages, such that a different number of topics needs to be modeled in each separate estimation, and the independent estimation may obscure similar patterns that are present in multiple languages but modeled only in some of them. Additional problems may arise where bi- or multi-lingual interpreters are not available and each language-specific topic model is interpreted by a different native speaker. While separate, language-specific topic models offer a fast and informative way to inductively identify highly prevalent themes (e.g., Heidenreich et al., 2019), differences identified through their comparison may express either valid differences or methodological artifacts. Unfortunately, there is no straightforward way to disentangle these possibilities.

Bridging Babel Using External Equivalence Information

If topic models are to be estimated jointly across multiple languages, the modeling algorithm must be enabled to map largely non-overlapping vocabularies into one common semantic space. To do so, computer scientists have developed so-called Multilingual Probabilistic Topic Models (MuPTM, Vulić et al., 2015; e.g., Boyd-Graber & Blei, 2009; Mimno et al., 2009), which rely on external resources to inform the model what words, expressions or documents are equivalent across languages (e.g., bilingual dictionaries or corpora with topically similar documents in different languages). The alignment of the languages takes place prior to the actual modeling process (Lind et al., 2019a): For example, the Polylingual Topic Model (Mimno et al., 2009) creates cross-lingual word clusters inferred from external resources such as Wikipedia pages in different languages covering the same topics (Lind et al., 2019a). More recently, Chan et al. (2020) have introduced *rectr* (Reproducible Extraction of Cross-lingual Topics using R), a technique which uses word embeddings obtained for different languages to map equivalent terms upon one another in a reproducible, automated fashion.

While MuPTM-approaches hold ample promise for multilingual text analysis, they are not free from limitations. Relying on external information to map words across languages, MuPTMs depend on the availability of suitable resources and are affected by any incommensurabilities therein (e.g., Wikipedia pages often differ across languages, especially in the context of politically sensitive issues; Massa & Scrinzi, 2012). Moreover, as MuPTMs explicitly remove any language-specific contents that cannot be mapped, the method is only suited to identify those meanings that are shared across linguistic boundaries. Accordingly, such approaches are valuable mostly for studying broad themes that can be generalized across languages, for example, general news beats or concurrent events (Chan et al., 2020; Vulić et al., 2015). However, they are limited in their capacity to identify comparative differences between different-language corpora. Finally, to date, only two software implementations of MuPT-Models are available: the Polylingual Topic Model (Mimno et al., 2009) in the standalone package MALLET by McCallum (2002), as well as Chan et al.'s (2020) *rectr* R library (see, Lind et al., 2019a).

Bridging Babel Using Translation

Machine translation (MT) methods can be considered a very intuitive way to solve the Babel problem: All source documents are translated into one common language. As a key advantage, the translation of multilingual corpora preserves any topical variation that may exist between and across languages. Usually, machine translation services such as Google Translate, DeepL, or Yandex are used for bulk translation (Reber, 2019).

At the same time, the reliance on machine translation can be criticized for several reasons. While the translation quality has improved significantly with the advances of machine learning technologies in the field of machine translation (Lotz & van Rensburg, 2014), not all translation attempts yield valid results. Even when they do, translation constitutes an invasive manipulation of the source data. All subsequent steps build on the assumption that the translated texts validly reflect not only the semantic contents of the original documents, but also their linguistic structuring. Consequently, it arguably matters which target language the material is translated into, and whether all languages are subjected to this procedure (Lucas et al., 2015). Critics also highlight that researchers cannot inspect the inner workings of the services or even save their underlying models, which limits reproducibility (Chan et al., 2020; Lind et al., 2019a). Moreover, while MT is much cheaper than human translation, it can also quickly exceed the research budget (e.g., Google Translate currently charges 20 USD per 1 million characters).

That being said, MT constitutes a pragmatic strategy that yields valuable insights and valid results (Lucas et al., 2015; De Vries et al., 2018) largely independent from the used service (Reber, 2019). Reber (2019) found that even translating the isolated words from the document-term matrix instead of complete documents still leads to comparable results, suggesting that the approach is reasonably robust. Another advantage is that it allows researchers to inspect the topics of a debate even if they do not understand the original language. However, in such cases, researchers' unfamiliarity with the linguistic and discursive uses and contexts of the original material may seriously compromise their ability to render valid interpretations, as well as their capacity to spot biases and errors in the translation process. The approach lends itself to comparative studies of textual contents where translations depend less on linguistic nuances or cultural context. Overall, the approach offers an accessible avenue for multilingual topic modeling, which can be applied to a wide variety of languages with limited effort.

Bridging Babel Using Multilingual Dictionaries

As compared to the aforementioned strategies, multilingual dictionaries (MD) represent a long-established and elaborated alternative approach. In a nutshell, they extend the tradition of researcher-constructed dictionaries for automated text analysis by operationalizing categories of interest through a comprehensive list of indicative expressions in multiple languages (Baden & Stalpouskaya, 2015; Lind et al., 2019b; Vliegenthart & Damstra, 2018). Conceptually, multilingual dictionaries can be viewed as a hybrid strategy: They deductively define what expressions in different languages map upon commensurable conceptual meanings, thus "translating" multilingual text into a common, abstract "language" on the semantic level. At the same time, this semantic "translation" is deliberately impoverished: Unlike machine translation, which aims to preserve any content in the original texts, dictionaries generally follow a deductive approach to text analysis, wherein researchers' operational judgment determines what contents are recorded, and how. Instead of attempting to map the full linguistic variation across languages, only those parts of the text that refer to a finite set of predefined conceptual meanings are retained, and all other information discarded (Doise et al., 1993; Popping, 2017). Moreover, the original text is further abstracted, merging different expressions, variants and sub-classes of recognized entities (e.g., a dictionary might regard "son," "sister," and "uncle" as equivalent references to an abstracted concept of "family members"). A text corpus coded with a multilingual dictionary represents the analyzed documents as sequences of semantic entities (Popping, 2017). Through the construction of the dictionary, the researcher fully controls the definition of the entities and equivalent expressions across different languages (Baden & Stalpouskaya, 2015). Once constructed, multilingual dictionaries not only permit the reconstruction of every stage of the research process, but they can also be applied free of scale without additional costs.

That said, multilingual dictionaries constitute a costly and effortful research instrument to construct. Both their reliance on deductive categories and their operationalization in multiple languages raises considerable demands on prior knowledge about the studied languages and topical discourses (Popping, 2017). Especially the definition of disambiguation rules that ensure that equivalent contents are captured in different languages requires elaborate validation (Lind et al., 2019b) and can severely impact the comparability of findings, as any remaining errors are systematic. Moreover, the strategy is structurally limited to modeling those contents whose relevance was anticipated by the dictionary, as all other textual contents are not retained.

Strengths and Limitations of the Four Approaches

In the comparison of all available approaches, the post-estimation bridging of separately calculated topic models is fraught with numerous limitations and cannot be recommended as a strategy for rigorous multilingual text analysis. MuPTMs are computationally advanced, but somewhat limited in their application, as their removal of language-specific contents focuses the estimation solely on those topical structures shared across linguistic boundaries. Machine translation (MT) and the use of multilingual dictionaries (MD) are intuitive, easy-to-implement techniques that rely on well-established algorithms and are capable of modeling both shared and language-specific topical structures.

At the same time, both approaches follow different roads in their treatment of multilingual text corpora. MT's inductive approach and flexibility contrasts against MD's reliance on deductive categories focusing and pre-structuring the analysis (Baden & Stalpouskaya, 2015). Where MT preserves the full content and nuance of natural language, MD merges equivalent expressions into abstracted semantic units, increasing focus by substantially diminishing the richness of the data. While topic modeling following MT can draw upon a wealth of tokens obtained from the (translated) texts, MD may be suitable for subsequent topic modeling only if a sufficient number of semantic entities is recognized in the analyzed text. Owing to the high demand on complex dictionaries, to the best of our knowledge, the MD-approach has not before been applied to the topic modeling of multilingual text corpora. In the following, we will introduce our test case and explain in detail how both machine translation and dictionary coding were implemented to enable the topic modeling of multilingual text corpora.

Study Design

For our study, we chose a challenging application case: We analyze two corpora, a Twitter corpus and a news corpus, both referring to the variegated conflicts between Israeli settlements and Palestinian Bedouin villages within "Area C," that is, the part of the West Bank that is under Israeli civil and military control. While the specific locale of the conflict shifts, there has been an ongoing public debate in both social media and professional news, which involves residents, activists as well as journalistic, political, and military actors (Schejter & Tirosh, 2012). Importantly, the debate takes place simultaneously in both local languages – Hebrew and Arabic – as well as English, owing to the considerable involvement of international activists in the conflict and the important role of global audiences and English-language media in its coverage. At the same time, each language community discusses the same conflict in somewhat different ways, focusing on different actors and assuming different political stances (e.g., Massa & Scrinzi, 2012). As a consequence, we expect to find both topics shared across all three languages, and topics that are notably more prevalent in, or even exclusive to, one of the three studied languages.

From a methodological standpoint, English serves as our reference case to connect to the primarily English-language focus of most computational text analysis. By contrast, both Hebrew and Arabic are Semitic, morphologically rich languages that use different non-Latin alphabets and pose a wide range of difficult challenges for natural language processing (Tsarfaty et al., 2013), including machine translation (Lucas et al., 2015). Of our two constructed test corpora, one is composed of news coverage in each language, capturing a relatively regular form of discourse structured by professional journalistic practices. The other corpus, which comprises social media posts, is rich in creative, colloquial and nonstandard language use and can be considered a hard case for our application (Baden, Kligler-Vilenchik & Yarchi, 2020).

For our study, both corpora are subjected to both machine translation and coded by the multilingual dictionary before the outcomes of each procedure are fed into a topic model. Based on the discussions above, we expect the MD procedure to generally yield more topics focused on the main dimensions of the political and violent conflict, which was the dictionary's primary focus. Due to the semantic abstraction achieved by the dictionary, we furthermore expect topics to contain conceptually well-defined terms that are unambiguous, and comparatively sparse in uninformative and/or off-topic words. In return, the MD should largely miss any additional patterns also present in the data that only touch upon conflict but involve other thematic domains that were disregarded by the dictionary. By contrast, the MT procedure should pick up all prevalent patterns regardless of their thematic focus,

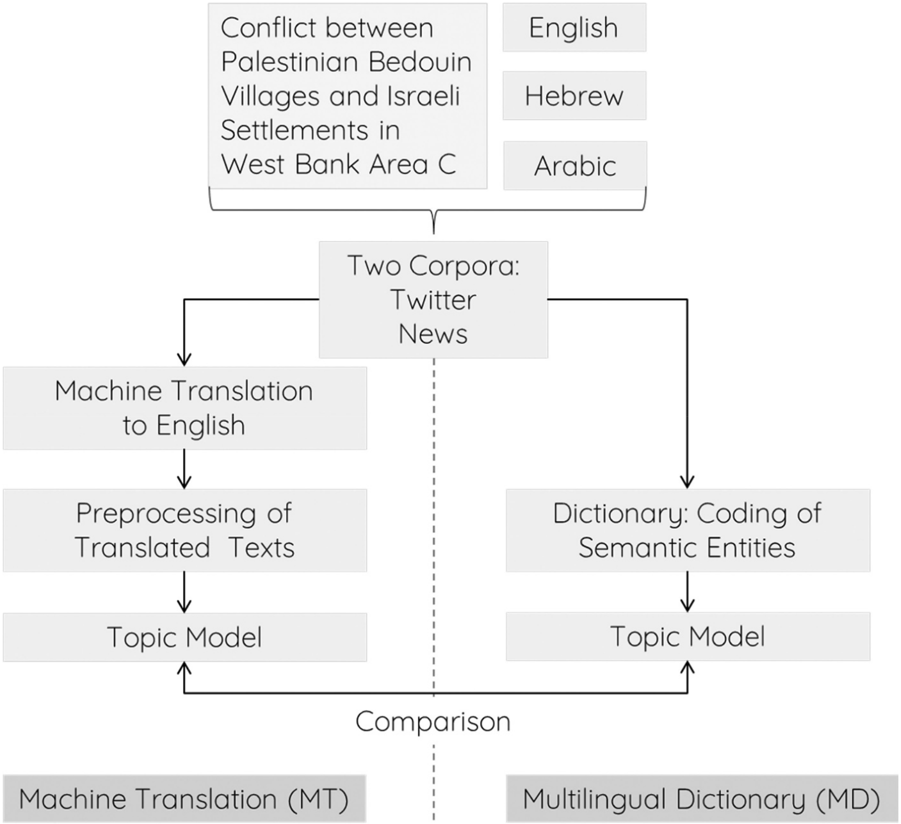
Conflict between Palestinian Bedouin Villages and Israeli Settlements in West Bank Area C
English
Hebrew
Arabic
Two Corpora: Twitter News
Machine Translation to English
Preprocessing of Translated Texts
Topic Model
Dictionary: Coding of Semantic Entities
Topic Model
Comparison
Machine Translation (MT)
Multilingual Dictionary (MD)

Figure 1. Study design.

likely including a higher share of off-topic or uninterpretable topics, as well as a stronger presence of uninformative or ambiguous words.[3]

We furthermore expect both approaches to perform relatively similarly with regard to the news corpus, where professional journalistic norms and routines limit both the topical diversity and the variability of language use. For the study of such relatively orderly discourse, hence, neither the dictionary's limited sensitivity to unconventional language nor its added topical focus should make a big difference. By contrast, we anticipate substantial differences in the topic models obtained from the social media data. Not only should the translation of social media corpus prove more error-prone; but we also expect the social media corpus to contain higher shares of unanticipated meanings and expressions that are not recorded by the dictionary. Figure 1 summarizes the design of the study.

Data and Methods

Data

Twitter Corpus

For our study, we include tweets posted between January 1, 2017 and December 31, 2018, written by users located within Israel and the Palestinian territories. We included tweets as relevant if they contained either a) at least one reference each to any pair of adjacent, rival Bedouin and Jewish settlements, b) at least one reference to a contested Bedouin village together with a reference to Jews or Israel or in Hebrew language; c) at least one reference to a contested Israeli settlement, together with a reference to Bedouins, Palestinians or in Arabic language; or d) at least one reference to either kind of place together with at least one manifest reference to the specific forms of violence (demolitions, protests, attacks etc.) involved in this conflict. Our full search string contained 223 names of Bedouin villages, 129 Israeli settlements and towns, 21 group references, and 16 conflict-related expressions in each language. The search string yielded just under 10,000 possible term combinations and was implemented using the social media analytics platform Crimson Hexagon. A human-readable version of the search string is provided in the Online-Appendix.[4] The query returned a total of 15,638 tweets in Hebrew, 56,007 in Arabic and 10,272 in English.[5] To reduce the imbalance between languages, a random sample of 15,000 Arabic tweets was drawn. After deduplication, the corpus retained a total of 34,973 unique tweets (Hebrew: 12,379, Arabic: 13,478, English: 9,116), which were minimally preprocessed (i.e., we removed emojis and URLs). Further text preprocessing steps were conducted separately for the two methodological tracks (see below).

News Corpus

For our News corpus, we obtained all coverage referring to the same conflict, using the same keywords and inclusion criteria laid out above, in three Israeli online news outlets published during the same period in Hebrew (*Ynet*), English (*Ynet News*) and Arabic (*Panet*). Each news outlet is a leading source of current news in its respective language in Israel, catering primarily to the Hebrew-speaking Jewish majority population, the sizable community of expatriates and recent immigrants, and the Arabic-speaking Palestinian minority within Israel, respectively. All three outlets belong to legacy news organizations (*Yedioth Ahronot* and *Panorama*) with a commitment to professional journalistic norms, a centrist stance and strong commercial orientation. They are produced by separate newsrooms, providing a 24/7 coverage of both factual news reports and evaluative commentary. Each outlet covers the conflict in the Israeli-controlled parts of the West Bank relying on specialized correspondents who, while to some extent reflecting their primary audiences' different allegiances in the conflict, strive to provide balanced and fact-based news (Caspi, 2008). All coverage was obtained directly from the news web pages, yielding a total of 14,526 news items from *Ynet*, 1,479 from *Ynet News*, and 7,761 from *Panet*. To reduce the imbalance in corpus sizes, a random sample of 2,000 news items was drawn for the Hebrew (*Ynet*) and the Arabic (*Panet*) language part of the corpus, resulting in a final corpus size of $N = 5{,}479$ documents (Hebrew: 2,000, English: 1,479, Arabic: 2,000).

Methods

Machine Translation

For the machine translation approach, the *Google Translate* API (Application Programming Interface) was leveraged. It enables the automated translation of chunks of text between thousands of language pairs. In our case, all tweets and all news items in Arabic and Hebrew were translated into English by passing the texts along to the API using *googleLanguageR* (Edmondson, 2018). After translation, documents were stripped from stop words, numbers, punctuation, separators, hyphens and symbols, and subjected to lemmatization. Accordingly, the Tweet:

> “*אלון בן דוד כותב היום במעריב מה שאני טוען שנים צריך לגרש את משפחות המחבלים לעזה כל הפיגועים יפסקו אחרי גירוש משפחה ראשונה*
> *איפוה ח״כ מהימין ישר לחוקק*”

is translated[6] and returned after preprocessing like this:

> “*alon ben david write today maariv say year should expel terrorist family gaza all attack cease after first family expulsion*”.

Finally, we applied relative pruning, i.e., we stripped words that occurred very rarely (in less than 0.5% of all documents) and extremely frequently (in more than 99.5% of all documents).[7]

Multilingual Dictionary

For the dictionary-based analysis, we used a large multilingual dictionary that had been developed within an international collaborative research project focusing on mediated conflict discourse (INFOCORE, www.infocore.eu). The dictionary, which is available upon request from the project consortium, was designed inductively based on a large-scale comparative analysis of conflict-related news, social media discourse, and political discourse in eight languages, including Hebrew, Arabic, and English. Specifically, the dictionary lists 3,738 unique semantic entities, which were obtained by a) identifying any meaning-carrying expressions via qualitative discourse analysis, b) grouping expressions used interchangeably, and c) translating them back and forth between languages to determine a segmentation that can be applied in equivalent ways in each language (Baden & Stalpouskaya, 2015). In this way, the dictionary recognizes a wide range of conflict-related activities (e.g., demand, destroy, de-escalate) and qualities (e.g., swift, secure, sincere), actors (e.g., Palestinian Authority, paramilitary, priest), objects (e.g., forgery, firearm, flag) and abstracts (e.g., information, independence, indictment), as well as places (e.g., Temple Mount, Tel Aviv, territories), times/events (e.g., colonial era, Christmas, Cave of Patriarchs massacre) and conflictual issues (e.g., Right to return, religious sites, rocket fire), each operationalized using a variety of direct and indirect expressions (e.g., the concept "hand-held guns" includes pistol, small arms, AK-47, rifle and several other related expressions) in each relevant language. The INFOCORE consortium specifically created the dictionary such that in any conflict-related news report, social media post, political speech or press release, most meaning-carrying entities would be classified, such that one can reconstruct the topical content and main claims included in each text from the coded references. Applying this dictionary, the tweet quoted above contains the following concepts:

> Written Note; Maariv; Today; Claim/Say; Past; Desirable/Should; Deportation; Terrorists; Family; Gaza Strip; Attack; Stop; Initial; Family; Deportation

After applying the dictionary the tweet would be represented by codes, reading:

> "*10783 30347 20412 10024 20402 10529 10327 30026 30223 40357 10340 20416 10384 30223 10327*"

To the already existing semantic entities/codes of the dictionary, we added the specific locales involved in the conflict studied here (Bedouin villages, Israeli settlements), obtaining a total of 4,081 unique concepts. Each entity is operationalized by a list of common, indicative keywords (including spelling variants) in each language, disambiguated based on their context, totaling more than 90,000 more or less complex search strings per language. The dictionary was repeatedly validated and improved until it achieved precision and recall scores well in excess of .75 in each language (see Baden & Stalpouskaya, 2015; Tenenboim-Weinblatt & Baden, 2021 for more detail). As the dictionary uses structural features of the documents (e.g., word distances, syntactic structure) for concept identification and disambiguation,[8] the dictionary was applied to the unprocessed text. As an output, this procedure represents each document as a sequence of five-digit codes indicating any concepts recognized in the text.

Topic Modeling

Owing to the brevity of tweets, which contain a maximum of 280 characters, our social media corpus required further adaptation prior to topic modeling. The application of topic models builds on the identification of coherently recurring word patterns (topics) within documents (Blei, 2012); short documents, however, often feature an insufficient frequency of word co-occurrences so that the topic modeling algorithm is unable to detect coherent patterns. This so-called "sparsity problem" leads to low quality topic models (Hong & Davison, 2010; Li et al., 2019). While a growing branch of research is concerned with developing new specialized models for collections of short texts (see e.g., Li et al., 2019),[9] a simple and established approach to address the sparsity-problem is to concatenate several short documents to a longer one (Guo et al., 2016). While this artificial document concatenation distorts the structure of the original data, the method leads to a significant increase of topic coherence in the resulting models (Steinskog et al., 2017). Following this approach, we temporarily concatenated sequences of up to five tweets authored by the same user in chronological order and generated the topic model based on these artificially prolonged documents. The topic model was then applied to the original Twitter corpus.

For the News corpus, such a procedure was deemed unnecessary as the news articles already feature a sufficient document length.[10] Accordingly, no separation between generation and application of the topic model was required.

For both methodological approaches applied to each corpus, we calculated Structural Topic Models (STM, Roberts et al., 2019). Structural Topic Models are well suited to be combined with multilingual corpora as they allow to control the prevalence and content of topics for covariate influence of the original language. In each case, multiple models were calculated, varying the number of topics (K= {10, 15, 20, 25, 30}).

Validation and Comparative Evaluation

Following the modeling stage, the obtained models were validated and compared using both quantitative and qualitative procedures.

For the quantitative comparison, we calculated the extent to which different topic models obtained from the respective methodological approaches break down the variation included in each corpus in similar ways. To do so, we use the Correlation Matrix Distance (CMD), a distance measure designed to determine the (dis)similarity between two quadratic correlation matrices (Herdin et al., 2005; Motta & Baden, 2013). Theoretically, the CMD ranges from one (no association between both matrices) to zero (both correlation matrices are identical, up to a scale factor).

For this estimation, we depart from the compared topic models' $n \times k$-sized document-topic (θ) matrices, wherein each of the $n = 1, \ldots, N$ rows correspond to a document in the respective corpus, and each of the $k = 1, \ldots, K$ columns to a topic in the respective topic model. In these matrices, the cell entries θ_{nk} represent the probability of document n to contain a given topic k – usually interpreted as the topic proportion that the respective document contains. Since the identity of the compared topics is uninformative for the comparison (i.e., one can permutate the order of topics (columns) in each matrix without changing the structural similarity), we transform the raw θ matrices into quadratic correlation matrices by calculating the dot product of θ with its transpose θ^T. As a result, we obtain, for each topic model, an $n \times n$ matrix C that reflects the extent to which different documents are composed of the same topics. The CMD then computes the distance between each pair of matrices C, which provides a direct measure of the extent to which the topics obtained by two compared topic models are distributed in (dis)similar ways over the same set of documents. Since the dimensionality of C depends solely on the number of the n documents, this transformation allows us to compare the θ's of different topic models independently of their number of topics K. Formula 1 defines the CMD for the two correlation matrices $C_{MT}$and C_{MD} that result from the two methods under investigation, as denoted by their subscripts.

$$CMD(C_{MT}, C_{MD}) = 1 - \frac{tr\{C_{MT} \cdot C_{MD}\}}{||C_{MT}|| \cdot ||C_{MD}||} \tag{1}$$

For the qualitative comparison, we first selected one best-fitting model per corpus and approach. Among all estimated topic models, we focused on those five models that offered the best fit based on the evaluation metrics offered by the STM package ($K = \{10, 15, 20, 25, 30\}$ for both corpora and both methods). Two of the authors jointly judged these models' interpretability, inspecting their top associated words (MT approach) and concepts (MD approach). Based on this information, we selected for the News corpus those topic models obtained for K = 25 (independently for both MT and MD), and for the Twitter corpus those models obtained for K = 30, for further validation. For these selected models, all topics were then carefully labeled and validated by the same two researchers. To ascertain the validity of topics and their labels, we selected for every topic a random sample of ten documents with a topic probability above a threshold of $t = 0.3$. Given that only few topics are prevalent in a single document, the threshold can be considered sufficiently high for a topic to be the most prevalent in the respective documents (see also Maier et al., 2018). Topics were confirmed as interpretable and labeled if, through a discursive process, both judges agreed on the same interpretation, which had to be supported both by the identified top words/concepts and by the inspected documents, read against our familiarity with the conflict and the referenced events.

Finally, topics were compared and matched between the two approaches applied to each corpus. Specifically, we considered topics (from different models) as matching if they expressed equivalent semantic meaning, based on the interpretation of associated top-words (MT) and top-codes (MD) and the validation of attributed meanings based on associated documents.[11] In this way, topics that were identified by both methodological approaches were distinguished from topics unique to one approach, and similar topics were inspected for remaining, meaningful differences. In the following, we report those patterns identified in the quantitative and qualitative comparison.

Results

Figure 2 shows the CMD values between each MT-MD pair of topic models (Figure 2a: Twitter corpus; Figure 2b: News corpus). Observed differences between both methodological approaches are notably larger for the News corpus than for the Twitter corpus. Contrary to expectations, thus, the greater variability of social media language did not diminish the overlap between the results obtained by the different approaches. For the Twitter corpus, the highest agreement was recorded between the $K = 25$ MT model and the $K = 20$ MD model, with agreement diminishing notably for smaller K values. For the News corpus, agreement was the highest for the $K = 15$ MT model and the $K = 10$ MD model, reflecting the slightly lesser thematic diversity of news contents. In both cases, differences tend to be smaller if K is equal or slightly smaller for MD than for MT, indicating MT-based topic models' consistent tendency to pick up a few patterns beyond those registered by the MD-based topic models. At the same time, the residual difference between even the most similar models documents that important, systematic differences remain between both approaches.

In the comparison between those respective topic models judged most interpretable in the qualitative validation, we investigated in which ways those topics picked up by the two approaches differ from one another (News corpus: $K = 25$; Twitter corpus: $K = 30$). To begin, for each corpus, less than half of the topics were identified by both approaches as nearly identical or equivalent (10/25 in the News corpus, 10/30 in the Twitter corpus). For the News corpus (10/25), but not for the Twitter corpus (4/30), we furthermore found a substantial share of topics that were partially matched between both models, i.e., one models' topic suggests a different focus than the other models' topic, but the thematic core of both is equivalent (see Endnote 11). As expected, the ratio of interpretable topics was notably higher for the News corpus (23/25 for MT, 23/25 for MD) than for the Twitter corpus, and lowest for the MD approach applied to the Twitter data (23/30 for MT, 18/30 for MD).

That said, almost all interpretable topics extracted from the Twitter corpus by the MD approach were in fact relevant to the overall discourse on violent conflict, well-delineated, distinct (i.e., they had little or no overlap with coherent word patterns of other topics), and easy to interpret (i.e., their top concepts allowed the researchers to understand what documents containing the topic were about). By comparison, the topics generated by the MT approach were less clearly delineated and less informative, and they tended to include a higher share of words whose role in the topic was ambiguous. Also, the MT approach picked up a higher share of topics that were interpretable but irrelevant to the studied conflict (4/30 for MT, 1/30 for MD). For the News corpus, both approaches identified a substantial share of topics unrelated to the conflict (7/25 for MT, 6/25 for MD), most of which were at least partly matched between both approaches. In both approaches, all but two topics identified within the News corpus were interpretable. Looking at the distribution of topics over the studied languages included in either corpus, the MD approach generated a substantial share of topics that

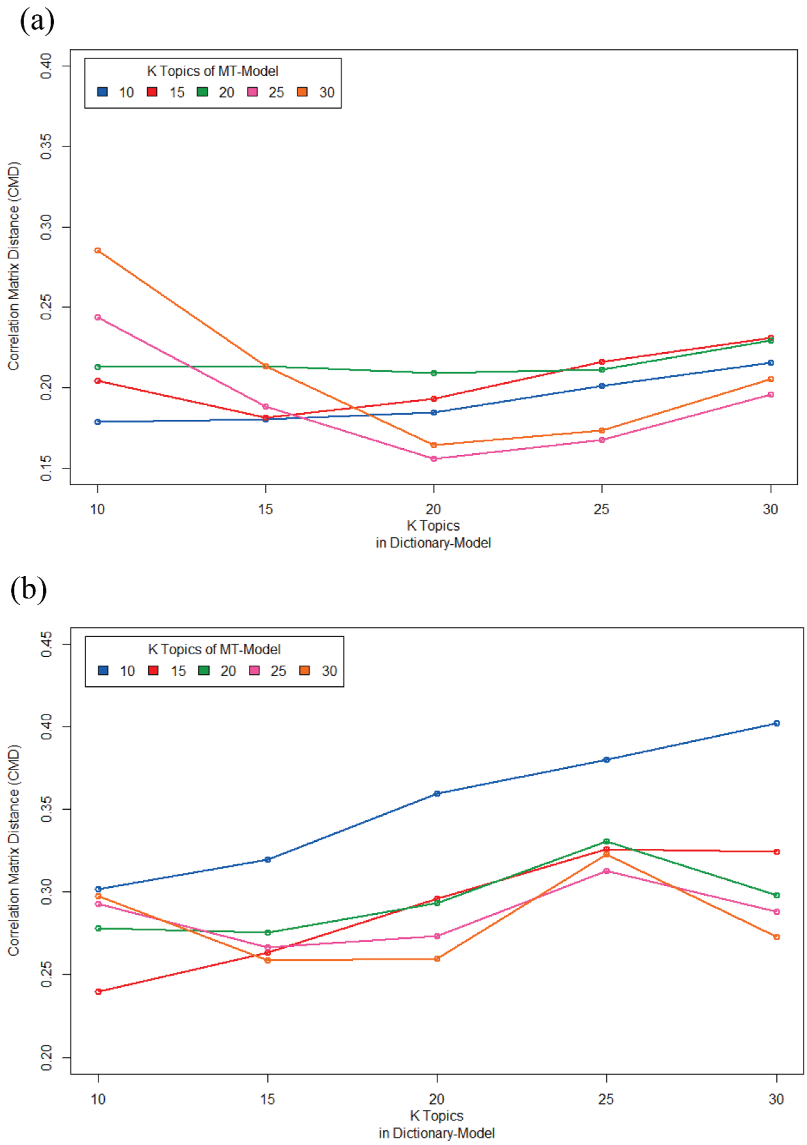

Figure 2. Correlation Matrix Distances of MT and MD Topic Models. *Note*. Panel A shows the CMD values for the Twitter Corpus; Panel B for the News Corpus.

originated almost exclusively (>90%) from one language alone. This tendency was notably less pronounced for the MT approach, especially with regard to the News corpus.

Concerning the semantic content of the identified topics, both approaches largely agreed (i.e., generated matching or partially matching topics) for those issues at the core of the investigated conflicts between Bedouins and Jewish settlements, and the Israeli-Palestinian conflict in general (for a summary of the models, see *Appendix B;* for the full depiction, please refer to the Online Appendix). In the News corpus, no clear pattern emerged for the differences between partially and unmatched topics obtained by either approach: In some cases, the MD topic model yielded the clearer and more relevant variant (e.g., “interfaith relations” vs. the broader “religious discourse” in the MT topic model; “protest against settlements” vs. “clashes in the West Bank”); and in other cases, it was the other way round (e.g., the MT topic “minority rights” vs. the broader MD topic “interethnic relations”). Both models picked up unique topics unrelated to the conflict (e.g., “food & recreation” for MT, “tourism” for MD) as well as some unique relevant, but rather broad topics (“violent conflict”). While there are noted differences between both models, both appear valid and exhibit no striking differences related to the chosen approach.

By contrast, there were systematic differences between the topics modeled for the Twitter corpus. Again, most topics at the core of the conflict were matched between both approaches; however, the MT approach included several broad, contextual topics that were missed by the MD approach, including both directly conflict-related ones (e.g., “human rights”) and topics without immediate relevance to the research focus (e.g., “spiritual”). The MD approach did not include most of the irrelevant topics. Those topics modeled by the MD approach, but missed by the MT approach, included relatively specific, well-delineated, and relevant sub-issues within the conflict (e.g., “dispute over agricultural land,” “violence by settlers”). Overall, the topics created by the MD approach tended to offer more nuance and specificity, while those created by the MT approach tended to be associated with broader and more general semantics.

Throughout both corpora, but especially for the Twitter corpus, topics were often easier to “read” and validate for the MT approach, which returns its top words as natural language words and lemmas; by comparison, the more abstract concept codes presented by the MD approach often required additional effort to understand which more specific expressions and instances in the original texts were part of a topic, resulting in a more difficult validation. Beyond the inability of the MD approach to model any topics composed of expressions not included in the dictionary, this difficulty at validation may account at least in part for the much longer list of uninterpretable topics obtained from the approach. Inversely, validation also revealed important limitations related to the MT approach. Specifically, we identified numerous false or inconsistent translations (e.g., many names and places were sometimes transliterated and sometimes translated), adding noise. In a non-negligible share of cases, translations were truncated or incomplete, cutting off sentences or simply retaining the original, untranslated text for part of the document. While less likely to result in major systematic errors, these issues limit the robustness and sensitivity of obtained topic models.

Discussion

Similarities and Differences between the Two Approaches

Overall, our analysis shows that both approaches deliver a valid, reasonably similar topical structuring of the two corpora. While important differences exist, both approaches perform well. Many of the noted differences in performance and topic identification can be traced to known strengths and limitations of each model, as well as the properties of the analyzed discourse. Differences were limited as long as the corpus was dominated by rather regular, consistent language use and terminology, as was the case for the news corpus (Van Dijk, 1985): As journalists use relatively few different words to refer to the same concepts, the dictionary’s capacity to merge conceptually equivalent terms did not substantially improve the model’s interpretability compared to the MT models; nor did the capacity to treat additional linguistic variation in the MT approach substantially outperform the sensitivity of the dictionary, which successfully recognized more or less the same expressions and patterns. This is true even for a variety of topics unrelated to the focus of the conflict, which are relatively common in the news (compared to the brief, typically monothematic tweets): Owing to the regular language of the news, the dictionary still validly recorded topics beyond its designed focus – only in diminished detail.

Both approaches’ respective strengths and limitations became apparent when they were applied to the social media corpus. On the one hand, MT’s sensitivity to expressions unrelated to the focal topic (and thus not recorded by the dictionary) enabled the MT model to construct a few topics that the MD model either missed entirely, or was unable to shape into interpretable form. On the other hand, the rich collection of equivalent or related expressions included in the dictionary, as well as its narrower focus on conflict-related expressions enabled the MD model to present concise and well-delineated topics. By contrast, the MT model responded to the greater variability of language by constructing more vague and less clearly delineated topics. By merging multiple expressions into abstract concepts, MD increased the amount of information available from the presented top words, at the cost of complicating validation, as these concepts were more distant from the original text. In this interpretation, the large number of uninterpretable topics in the MD model for Twitter reflects a sparseness issue (Hong & Davison, 2010) created by the dictionary's incomplete recognition of unanticipated, off-topic semantic patterns. While it may thus be preferable to carefully filter out off-topic content prior to applying the dictionary, MT remains unaffected by the distinction between on- and off-topic content.

Contrary to our expectations, the greater variability of language on social media did not lead to an increased quantitative distance between the topic models obtained from the two approaches. In fact, the distance was diminished for the Twitter corpus. One possible explanation is that in spite of the reduced variability of journalistic language, the greater length of journalistic texts offers more space for off-topic content. Less restrictive document lengths not only facilitate opportunities for journalists to introduce additional topical foci (e.g., background knowledge and individual stories beyond the shared conflict focus), they also increase the chance that, through search string queries, articles are captured for analysis that mention conflict-related sites and events only in passing (e.g., weather reports). Accordingly, the linguistic variability of the news corpus would be greater than assumed. By contrast, the brevity of tweets ensured that the captured texts are largely "on topic." In this vein, the high, but relatively unsystematic linguistic variability of the social media corpus led to more similar topic models, while the limited, but more systematic linguistic variability of the news corpus resulted in more different models.[12] Additional research will be needed to fully understand this unexpected finding. Both approaches are mildly affected by errors in the harmonization across languages – be that translation errors in the MT approach (Lotz & van Rensburg, 2014), or terms that were insufficiently disambiguated in the dictionary (Baden & Stalpouskaya, 2015). While errors in the dictionary were arguably more consequential, most of these can be fixed with limited human effort; translation errors, by contrast, lie beyond our control. A somewhat more worrying finding concerns the MD models' pronounced tendency to create topics that are dominated by one language alone. While the MT models also include such topics, they present several valid and relevant topics that exist in similar shape in all three languages. Additional testing will be required to decide whether their separation by language in the MD models reflects valid variations between languages that were lost in the machine translation, or was caused by unintended differences in the dictionary's language-specific indicators.

Pragmatically, the MT approach is a fast, accessible, and inexpensive means to topic modeling of multilingual corpora (Lucas et al., 2015). By now, services such as Google Translate provide reasonable translations even for social media content, which is riddled by unconventional language and slang (Reber, 2019). While MT still created numerous mistranslations and regularly misconstrued the grammatical structure in our (admittedly difficult) translation of Hebrew and Arabic text, most documents were translated adequately. As the use of these black-boxed services does not allow for complete reproducibility (Chan et al., 2020), it may be wise to store and archive machine translation input and output data, and conduct post-hoc checks for systematic translation-related errors. Yet, the approach yielded valid, robust, and interpretable results for both corpora, and its ease of use and flexibility conceivably outweigh its limitations in many applications within and beyond academia. Specifically for investigations that aim to recognize topical regularities across languages in a fully inductive fashion, MT's inclusive, relatively agnostic treatment of the textual data clearly offers a superior solution.

By comparison, the added focus offered by the MD approach appears to apply mostly to the study of highly variable text (such as our Twitter corpus) but makes little difference for the analysis of regular language use (such as our News corpus). The approach offers tight control, ensuring the commensurability of recognized contents at a semantic level and enabling the researcher to focus the analysis in a theory-driven fashion (Popping, 2017). These advantages come at the expense of considerable effort required to construct and validate multilingual dictionaries, unless suitable dictionaries are available or can be easily adjusted (Lind et al., 2019b). Based on our findings, the added complexity of constructing multilingual dictionaries offers little added benefit when studying more regular forms of discourse, but may raise some consequential trade-offs when applied to more unruly data. In particular, multilingual dictionaries offer important advantages for focusing the analysis on only selected contents of a corpus, while ignoring others. Likewise, the possibility to define wider constructs that include a variety of expressions may be valuable to introduce some limited deductive control into the topic modeling procedure. The approach thus offers superior results specifically for "abductive" analyses (Reichertz, 2007) that include both deductive and inductive elements. In view of these differential strengths and limitations, summarized in Table 1, researchers may choose to prioritize the deductive control and focus afforded by MD, or prefer the inductive, more data-driven MT approach with its greater sensitivity to contextualizing topics, depending on their respective research interests.

At the same time, there may also be strategies that strike some balance between both approaches: For instance, Reber's (2019) experience with translating document-term matrices suggest that at least some parts of multilingual dictionaries may be valuably augmented (Kantner & Overbeck, 2020) or even constructed with the help of machine translation, decreasing development costs. Inversely, researchers might apply more restrictive, theoretically guided preprocessing strategies to exclude words unrelated to a pursued research question, thus trying to focus also MT-based topic models on specific contents only. In the pursuit of identifying valid solutions for the Babel problem, both approaches contribute complementary features that have the capacity to advance future development.

A More Informative Perspective on Validation

Beyond its primary focus on two methodological approaches to topic modeling multilingual corpora, our study suggests a somewhat different perspective upon the comparative validation of available methods (Grimmer & Stewart, 2013). Acknowledging that there cannot be a "correct" gold standard for the extraction of inductively defined topics, our study eschews a simple evaluation of which approach gets more classifications "right," in favor of a more qualitative understanding of methodological differences. Based on our investigation, we cannot say which approach is better, but instead focus on what each approach is better *at*.

Pragmatically speaking, our study can be understood as evidence that even much more elaborate strategies for ensuring cross-lingual commensurability do not systematically outperform the rather simple and accessible avenues offered by machine translation. Methodologically, it contributes to our understanding of the implications of concatenating algorithmic procedures (Baden et al., 2021). While the MD's representation of the original corpus tosses out much of the linguistic variation, this loss did not substantially compromise the sensitivity and validity of the subsequent topic modeling procedure, and even improved its interpretability. Inversely, the MT models' consistent capacity to pick up topics shared across linguistic boundaries attests that the primary topical organization of the corpus is well-

Table 1. Comparison of MT and MD.

	Machine Translation (MT)	Multilingual Dictionary (MD)
Cost/Effort	low, costs per word	high, but free once constructed
Researcher Control	only over processes prior or post translation (e.g., preprocessing, parameter settings of the topic model)	full control over construct definition and inclusion of contents
Sensitivity	to all contents	only to contents defined as relevant
Reproducibility	limited by dynamic black box MT services	perfect (deterministic)
Characteristics of Topics	broader, likely to include ambiguous, redundant or uninformative top words	more focused, information-rich top concepts
Ease of Validation	easy: top words occur in text	complex: need to bridge top concepts and words in text
Primary Application	inductive research; corpora that include relevant, unanticipated linguistic variation	abductive research; corpora wherein most relevant linguistic variation can be anticipated

preserved by the translation, even if translations occasionally fail in detail. This was true even for the rather demanding translation from two Semitic, morphologically rich languages using non-Latin scripts, in a discourse rich in slang and creative language use (Lucas et al., 2015; Tsarfaty et al., 2013). Both findings, furthermore, shed additional light on the robustness of topic models (Blei, 2012): Topic models uncover highly prevalent and systematic topical patterns in the organization of discourse largely unperturbed by even major transformations of the original data. Less prevalent and convoluted patterns, however, suffer severe robustness issues.

Additional research is needed to determine whether those specific strengths and weaknesses laid out above hold more generally. Specifically, it will be valuable to include also approaches based on external information, such as Chan et al.'s (2020) recently published *rectr* method, in the comparative validation of multilingual topic models' overall performance, and the identification of qualitative differences between them. Another intriguing strategy may be to consider keyword assisted topic models (Eshima et al., 2020) in combination with MT or MD. Further investigations may also reveal the influence of different balances between included languages (e.g., a few Spanish language tweets in a large US corpus) on multilingual topic models, and how these interact with different preprocessing and modeling procedures. Beyond our focus on the topic modeling of multilingual corpora, our investigation thus contributes to a better understanding of the ways in which topic models respond to different properties and manipulations of the underlying data, and a more qualitative understanding of the resulting differences.

Notes

1. A "gold standard" needed to determine model performance in absolute terms is unavailable for the inductive modeling of topical structures, as the same texts generally permit different thematic abstractions and topical orderings (Van Dijk, 1985).
2. Typically, this is the case for named entities and some words of foreign (e.g., Greek, Latin) origin; where languages use different scripts, as in the case of English, Hebrew, and Arabic, their vocabularies overlap only in their use of numbers.
3. Off-topics or uninterpretable topics are frequent phenomena raised by topic models, to the point that the most prevalent topics of a model may be so-called background-topics (Maier et al., 2018). These topics consist of terms that which occur most frequently across the whole corpus but are not specific to any given, semantically valid topic. Uninformative and ambiguous words are also common within any modeled topic. Without extensive prior cleaning, topics are frequently populated with words that are not completely coherent with the concept that the rest of the topic represents. For an overview, see Mimno et al. (2011).
4. While linguistic norms differ widely between Twitter and the news, place names – which made up the lion's share of our keywords – are among the least variable words in either kind of text; the main discrepancies between the thematic focus of both corpora thus are likely to arise from nonconventional or pejorative references to either side in the conflict, or slang references to specific conflict actions that would be used on Twitter but not in the news. The full search string can be found in the Online Appendix: https://bit.ly/3pFN8jw
5. While most documents are written in only one of the languages, few documents (tweets) mix languages (so-called code-switching). In the MT approach, English contents in Hebrew/Arabic texts would cease to stand out post-translation, and enter the modeling as any other token; contents in other different languages would remain untranslated and enter the modeling as noise. The MD approach, by contrast, ignores any contents expressed in a language other than the one expected.
6. "*Alon Ben David writes today in Ma'ariv what I have been saying for years should expel the terrorists' families to Gaza. All attacks will cease after the first family expulsion.*"
7. For the removal of stop words, numbers, separators, hyphens and symbols we used the built-in functionality of quanteda's tokens-function (Benoit et al., 2018). For lemmatization we applied a lemmatizer of the textstem package (Rinker, 2018) and for relative pruning a custom function written by Wiedemann and Niekler (2017).
8. For example, the term "general" was disambiguated between 10654 [General/Overall] and 30019 [Soldiers: Senior Officers] by examining whether other military-related terms were found within 20 words' distance, counting syntactic boundaries as multiple words; see Baden and Stalpouskaya (2015).
9. See e.g., Twitter-LDA (Zhao et al., 2011) or the Relational Biterm Topic Model (Li et al., 2019).
10. For the document length statistics, please see Appendix A.
11. To match topics, we considered both the top words and concepts identified by the MT/MD models, and the semantic interpretation obtained through the previous procedure of ascertaining topic validity and topic labels.

In order to be considered a matched pair of topics, they had to fulfill one of the following two criteria: 1) Top words of an MT model match directly with top concepts of the MD model and vice versa (i.e., MT top words are identical or equivalent to MD concepts); 2) words/concepts add information that is absent in the other topic but is consistent with the interpretation. If words/concepts add information that would privilege a different interpretation while some of the top words and concepts directly matched, topics were considered partly matched (e.g., a shared focus on violence and Jewish settlements, once with a focus on settlers and once with a focus on protesters and police).

12. In addition, unlike the qualitative comparison, the quantitative comparison responded not only to salient patterns in the textual content that were constitutive of topics' interpretation, but also to any differences in the distribution of rather uninformative contents over the identified topics, possibly inflating the CMD measure.

Acknowledgments

We would like to thank the Center for International Cooperation at the Free University of Berlin fortheir seed funding of the fruitful partnership with the HebrewUniversity of Jerusalem. Additionally, we thank Barbara Pfetsch for her generous financing to supply Open Access to this article.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [290045248 – SFB 1265].

Notes on contributors

Daniel Maier (Dr. phil., Free University of Berlin) is a data scientist at the German Cancer Consortium in Heidelberg. During the work on this article, he was a postdoctoral researcher at the Free University of Berlin and the Collaborative Research Center “Refiguration of Spaces” (CRC 1265) funded by the German Research Foundation (DFG). His research interests include various techniques of content analysis, network analysis, social media communication, and public health.

Christian Baden (PhD, University of Amsterdam) is associate professor at the Department of Communication & Journalism, Hebrew University of Jerusalem. His work focuses on the negotiation and evolution of contested and shared meanings in dynamic public discourse, advancing theory and methodology in textual analysis and the study of social communication.

Daniela Stoltenberg is a research associate at the Institute for Media and Communication Studies at Free University of Berlin and the Collaborative Research Center 1265 “Re-Figuration of Spaces.” Her research interests include digital public spheres, online social movements, urban communication, and computational communication science.

Maya de Vries-Kedem (PhD, Hebrew University of Jerusalem) is a lecturer at the Department of Communication & Journalism and at the Swiss Center for Conflict Resolution Studies at the Hebrew University of Jerusalem. Her research interests are digital activism, digital participation, the Israeli-Palestinian conflict and East Jerusalem.

Annie Waldherr (PhD, Free University of Berlin) is Professor of Computational Communication Science in the Department of Communication at the University of Vienna. She studies the changing structures and dynamics in today's digitized public spheres, combining computational and conventional empirical methods.

ORCID

Daniel Maier http://orcid.org/0000-0001-6266-8987
Christian Baden http://orcid.org/0000-0002-3771-3413
Daniela Stoltenberg http://orcid.org/0000-0001-9334-1514
Annie Waldherr http://orcid.org/0000-0001-7488-9138

References

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. Communication Methods & Measures, 14 (3), 165–183. doi:10.1080/19312458.2020.1803247

Baden, C., Pipal, C., Schoonvelde, M. & van der Velden, M. A. C. G. (2021). Three gaps in computational methods for social sciences: A research agenda. 71st ICA Annual Conference [Virtual Conference], Denver, CO

Baden, C., & Stalpouskaya, K. (2015). *Common methodological framework: Content analysis. A mixed-methods strategy for comparatively, diachronically analyzing conflict discourse*. Ludwig Maximilian University Munich:INFOCORE Working Paper 10/2015. https://www.infocore.eu/wp-content/uploads/2016/02/Methodological-Paper-MWG-CA_final.pdf

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Boyd-Graber, J., & Blei, D. M. (2009). Multilingual topic models for unaligned text. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 75–82). AUAI Press. Montreal, Canada.

Caspi, D. (2008). Israel: Media system. In W. Donsbach (Ed.), *The international encyclopedia of communication*. Wiley.

Castells, M. (2008). The new public sphere: Global civil society, communication networks, and global governance. *The Annals of the American Academy of Political and Social Science*, *616*(1), 78–93. https://doi.org/10.1177/0002716207311877

Chan, C.-H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., Van Atteveldt, W., & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, *14*(4), 285–305. https://doi.org/10.1080/19312458.2020.1812555

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Doise, W., Clemence, A., & Lorenzi-Cioldi, F. (1993). *The quantitative analysis of social representations*. Harvester Wheatsheaf.

Edmondson, M. (2018). *googleLanguageR: Call Google's 'Natural Language' API, 'Cloud Translation' API, 'Cloud Speech' API and 'Cloud Text-to-Speech' API*. (R package version 0.2.0) [Computer software]. https://CRAN.R-project.org/package=googleLanguageR

Eshima, S., Imai, K., & Sasaki, T. (2020). *Keyword assisted topic models*. arXiv. https://arxiv.org/abs/2004.05964

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101* (suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. https://doi.org/10.1177/1077699016639231

Heidenreich, T., Lind, F., Eberl, J. M., & Boomgaarden, H. G. (2019). Media framing dynamics of the 'european refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies*, *32*(Special Issue 1), i172–i182. https://doi.org/10.1093/jrs/fez025

Herdin, M., Czink, N., Ozcelik, H., & Bonek, E. (2005). Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels. *2005 IEEE 61st Vehicular Technology Conference* (Vol.1, pp. 136–140). IEEE.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the first workshop on social media analytics* (pp. 80–88). Association for Computing Machinery. New York, NY. https://doi.org/10.1145/1964858.1964870

Kantner, C., & Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETAWerkstatt* (pp. 169–189). De Gruyter. https://doi.org/10.1515/9783110693973-008

Kligler-Vilenchik, N., de Vries Kedem, M., Maier, D., & Stoltenberg, D. (2020). Mobilization vs. demobilization discourses on social media. *Political Communication*. 1–20. https://doi.org/10.1080/10584609.2020.1820648

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Social science. Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Li, X., Zhang, A., Li, C., Guo, L., Wang, W., & Ouyang, J. (2019). Relational biterm topic model: Short-text topic modeling using word embeddings. *The Computer Journal*, *62*(3), 359–372. https://doi.org/10.1093/comjnl/bxy037

Lind, F., Eberl, J. M., Galyga, S., Heidenreich, T., Boomgaarden, H. G., Jiménez, B. H., & Berganza, R. (2019a). *A bridge over the language gap: Topic modelling for text analyses across languages for country comparative research*. University of Vienna: Working Paper of the REMINDER-Project. https://www.understandfreemovement.eu/wp-content/uploads/2020/01/D8.7.pdf

Lind, F., Eberl, J. M., Heidenreich, T., & Boomgaarden, H. G. (2019b). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, *13*, 4000–4020.

Lotz, S., & van Rensburg, A. (2014). Translation technology explored: Has a three-year maturation period done Google Translate any good? *Stellenbosch Papers in Linguistics Plus*, *43*, 235–259. https://doi.org/10.5842/43-0-205

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Lück, J., Wessler, H., Wozniak, A., & Lycarião, D. (2018). Counterbalancing global media frames with nationally colored narratives: A comparative study of news narratives and news framing in the climate change coverage of five countries. *Journalism*, *19*(12), 1635–1656. https://doi.org/10.1177/1464884916680372

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Massa, P., & Scrinzi, F. (2012). Manypedia: Comparing language points of view of Wikipedia communities. *Proceedings of the eighth annual symposium on wikis and open collaboration* (pp. 1–9). Association for Computing Machinery. Linz, Austria. https://doi.org/10.1145/2462932.2462960

McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. [Computer software]. University of Massachussetts at Amherst, MA. http://www.cs.umass.edu/~mccallum/mallet

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 2* (pp. 880–889). Association for Computational Linguistics.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics. Edinburgh, United Kingdom.

Motta, G., & Baden, C. (2013). Evolutionary factor analysis of the dynamics of frames: Introducing a method for analyzing high-dimensional semantic data with time-changing structure. *Communication Methods and Measures*, *7*(1), 48–82. https://doi.org/10.1080/19312458.2012.760730

Popping, R. (2017). Online tools for content analysis. In N. G. Fielding, R. Lee, M., & G. Blank (Eds.), *The Sage handbook of online research methods* (2nd ed., pp. 329–343). Sage.

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Reichertz, J. (2007). Abduction: The logic of discovery of grounded theory. In A. Bryant & K. Charmaz (Eds.), *The Sage handbook of grounded theory* (pp. 214–228). Sage.

Rinker, T. W. (2018). *Textstem: Tools for stemming and lemmatizing text*. (R package, Version 0.1.4) [Computer software]. https://cran.r-project.org/web/packages/textstem/index.html

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Schejter, A. M., & Tirosh, N. (2012). Social media new and old in the Al-'Arakeeb conflict: A case study. *The Information Society*, *28*(5), 304–315. https://doi.org/10.1080/01972243.2012.708711

Steinskog, A., Therkelsen, J., & Gambäck, B. (2017). Twitter topic modeling by tweet aggregation. *Proceedings of the 21st Nordic conference on computational linguistics* (pp. 77–86).

Tenenboim-Weinblatt, K., & Baden, C. (2021). Gendered communication styles in the news: An algorithmic comparative study of conflict coverage. *Communication Research*,*48*(2), 233–256.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, *39*(1), 15–22. https://doi.org/10.1162/COLI_a_00133

van Dijk, T. A. (1985). tructures of news in the press. In T. A. van Dijk (Ed.), *Discourse and communication* (pp. 69–93). De Gruyter.

Vliegenthart, R., & Damstra, A. (2018). Parliamentary questions, newspaper coverage, and consumer confidence in times of crisis: A cross-national comparison. *Political Communication*, *36*(1), 17–35. https://doi.org/10.1080/10584609.2018.1478472

Volkmer, I. (2014). *The global public sphere: Public communication in the age of reflective interdependence*. Polity.

Vulić, I., De Smet, W., Tang, J., & Moens, M.-F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, *51*(1), 111–147. https://doi.org/10.1016/j.ipm.2014.08.003

Wiedemann, G., & Niekler, A. (2017). Hands-on: A five day text mining course for humanists and social scientists in R. *Proceedings of the 1st workshop on teaching NLP for digital humanities*. Berlin, Germany, Association for Computational Linguistics.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing Twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *European conference on information retrieval* (pp. 338–349). Springer. https://doi.org/10.1007/978-3-642-20161-5_34