

Multiword expressions in lexical resources

Linguistic, lexicographic, and
computational perspectives

Edited by

Voula Giouli

Verginica Barbu Mititelu

Phraseology and Multiword Expressions 6



Phraseology and Multiword Expressions

Series editors: Agata Savary (University of Tours, Blois, France), Manfred Sailer (Goethe University Frankfurt a. M., Germany), Yannick Parmentier (University of Lorraine, France), Victoria Rosén (University of Bergen, Norway), Mike Rosner (University of Malta, Malta).

In this series:

1. Manfred Sailer & Stella Markantonatou (eds.). Multiword expressions: Insights from a multilingual perspective.
2. Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.). Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop.
3. Yannick Parmentier & Jakub Waszczuk (eds.). Representation and parsing of multiword expressions: Current trends.
4. Schulte im Walde, Sabine & Eva Smolka (eds.). The role of constituents in multiword expressions: An interdisciplinary, cross-lingual perspective.
5. Trklja, Aleksandar & Łukasz Grabowski (eds.). Formulaic language: Theories and methods.
6. Giouli, Voula & Verginica Barbu Mititelu (eds.). Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives.

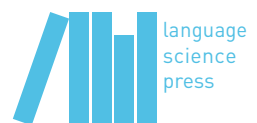
Multiword expressions in lexical resources

Linguistic, lexicographic, and
computational perspectives

Edited by

Voula Giouli

Verginica Barbu Mititelu



Voula Giouli & Verginica Barbu Mititelu (eds.). 2024. *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives* (Phraseology and Multiword Expressions 6). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/440>

© 2024, the authors

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/> 

ISBN: 978-3-96110-470-3 (Digital)

978-3-98554-099-0 (Hardcover)

ISSN: 2625-3127

DOI: 10.5281/zenodo.10949960

Source code available from www.github.com/langsci/440

Errata: paperhive.org/documents/remote?type=langsci&id=440

Cover and concept of design: Ulrike Harbort

Proofreading: Alexandr Rosen, Annika Schiefner, Brett Reynolds, Elen Le Foll, Eleni Koutsomitopoulou, Elliott Pearl, Harold Somers, Jean Nitzke, Jeroen van de Weijer, Ludger Paschen, Mike Rosner, Nathan Schneider, Rebecca Madlener, Sebastian Nordhoff

Fonts: Libertinus, Arimo, DejaVu Sans Mono

Typesetting software: $\text{X}_{\text{L}}\text{A}_{\text{T}}\text{E}_{\text{X}}$

Language Science Press

xHain

Grünberger Str. 16

10243 Berlin, Germany

<http://langsci-press.org>

Storage and cataloguing done by FU Berlin

Freie Universität  Berlin

Contents

Acknowledgments	iii
Preface	v
1 LEMUR: A lexicon of Czech multiword expressions Hana Skoumalová, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondříčka & Milena Hnátková	1
2 Description of Pomak within IDION: Challenges in the representation of verb multiword expressions Stella Markantonatou, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chiril, Dimitrios Karamatskos, Nicolaos Valeontis & George Pavlidis	39
3 A uniform multilingual approach to the description of multiword expressions Svetlozara Leseva, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu	73
4 Representation of multiword expressions in the Bulgarian integrated lexicon for language technology Petya Osenova & Kiril Simov	117
5 A FrameNet approach to deep semantics for MWEs Voula Giouli, Vera Pilitsidou & Hephestion Christopoulos	147
6 Multiword expressions, collocations and the OntoLex vocabulary Christian Chiarcos, Maxim Ionov, Elena-Simona Apostol, Katerina Gkirtzou, Besim Kabashi, Anas Fahad Khan & Ciprian-Octavian Truică	187
7 MWE-Finder: Querying for multiword expressions in large Dutch text corpora Jan Odijk, Martin Kroon, Sheean Spoel, Ben Bonfil & Tijmen Baarda	229

Contents

8	Collecting and investigating features of compositionality ratings	
	Sabine Schulte im Walde	269
9	Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results	
	Therese Lindström Tiedemann, David Alfter, Yousuf Ali Mohammed, Daniela Piipponen, Beatrice Silén & Elena Volodina	309
	Index	349

Acknowledgments

This volume would not have been possible without the help of the reviewers. They provided the authors with constructive feedback and us, the editors, with comments relevant to deciding upon the acceptance or rejection of the numerous contributions received.

- Archna Bhatia
- Lars Borin
- Mathieu Constant
- Thierry Declerck †
- Ismail el Maarouf
- Kilian Evang
- Tunga Gungor
- Kyo Kageura
- Ilan Kernerman
- Cvetana Krstev
- Tita Kyriakopoulou
- Svetlozara Leseva
- Timm Lichte
- Irina Lobzhanidze
- Stella Markantonatou
- Nurit Melnik
- Petya Osenova
- Viktor Pekar
- Alain Polguere
- Alexandre Rademaker
- Mike Rosner
- Nathan Schneider
- Sabine Schulte im Walde
- Gilles Serasset
- Ranka Stanković
- Manfred Stede
- Shiva Taslimipoor
- Beata Trawinski
- Veronika Vincze
- Nianwen Xue

We are also grateful to our editors-in-charge, Carlos Ramisch and Lonneke van der Plas, who provided their feedback and guidance where needed, as well as to Language Science Press representatives, Sebastian Nordhoff and Felix Kopecky, for the technical support they have offered us along the way.

Preface

Multiword Expressions (MWEs) have received growing attention from scholars in various disciplines, from theoretical to applied linguistics and psycholinguistics and from lexicography for human users to Human Language Technology. In this respect, linguists seek to account for their properties and to define typologies thereof; in applied linguistics, MWEs of various kinds pose issues for language learning and teaching; issues relative to the acquisition, and processing of MWEs, as well as the way they are stored in the mental lexicon constitute the focus of attention in psycholinguistic research, whereas lexicographers are well aware of the importance of their presence in dictionaries (Evert 2004) and strive to define optimal representation formats tailored to meet the needs of humans and machines alike. Computational linguists on the other hand are concerned with MWE processing, primarily with their identification and discovery in corpora, as well as with their cross-lingual equivalence, even though MWEs might be of importance in other downstream tasks too. Given the inherent idiosyncrasies of MWEs, all these tasks are considered problematic.

MWE identification and discovery are seen as the two facets of MWE processing (Constant et al. 2017) and lexical resources of all sorts remain at the heart of both: the former could be made easier given a resource lexicon containing them, while the latter could contribute to the enhancement of such a resource (Ramisch 2023). Consequently, Savary et al. (2019) proposed the deployment of MWE-related lexical resources as a possible solution for improving MWE processing; therefore, despite the ever-increasing effort to develop corpora of considerable size as well as language models of all kinds, MWE lexica are still needed.

An important open issue in the literature dedicated to this topic is the representation of MWEs in lexical resources. The time when mere lists of MWEs were considered lexicons has passed, and rich descriptions of MWEs are being created or enriched, with special attention paid to their idiosyncrasies at various linguistic levels (lexical, morphological, syntactic, and semantic).

This volume contains chapters that paint the current landscape of MWE representations in lexical resources from the perspectives of their robust identification and computational processing. Both large-size general lexica and smaller MWE-centred ones are included, with special focus on the representation decisions and

mechanisms that facilitate their usage in NLP tasks. The presentations go beyond the morpho-syntactic description of MWEs, into their semantics. These chapters confirm that no common technical solution to the problem of MWE lexical representation exists, as already pointed out in the literature (Lichte et al. 2019).

One challenge in representing MWEs in lexical resources is ensuring that the variability along with extra features required by the different types of MWEs can be captured efficiently. In this respect, recommendations for representing MWEs in mono- and multilingual computational lexicons have been proposed; these focus mainly on the syntactic and semantic properties of support verbs and noun compounds and their proper encoding (Calzolari et al. 2002, Copestake et al. 2002).

The interest in developing MWE lexicons results either in those that are MWE-dedicated (see the chapters authored by Skoumalová et al., Markantonatou et al. and Leseva et al.) or in those that are MWE-aware (see Osenova and Simov's contribution and Giouli et al.'s one). Though most of the time the focus is on a language's MWE system, there is also concern for language varieties (see Markantonatou et al.).

All chapters are circumscribed by the NLP domain, with the exception of Tiedemann et al.'s work in which language learning and teaching is the field of interest. The NLP-oriented chapters are concerned with facilitating the processing of texts containing MWEs, while the latter aims at improving learners' fluency by promoting a better understanding of MWE's degree of compositionality and properly handling this approach in teaching materials. However, compositionality, as a key characteristic of MWEs, is a challenge not only for machines, but also for human users, be they language learners, who are the target of Tiedemann et al.'s experiments, or native speakers, as reported in the chapter authored by Schulte im Walde.

There are languages for which language resources have been created over a long period and it is high time they were interconnected to better exploit their potential synergy. Osenova and Simov use the catena representation to this end, while Chiarcos et al. present a solution for standardized formatting of resources, namely the Linked (Open) Data paradigm, which can also help overcome resource scarcity of languages by complementing linguistic information in one resource with information from one or more other resources.

A resource such as WordNet (Miller 1995, Fellbaum 1998) has the advantage of encoding the meaning of MWEs in a relational manner: on the one hand, they participate in a synonymy relation at the level of synsets (MWEs may be part of a synset alongside either simple words or other MWEs); on the other hand, such synsets are themselves interlinked with other synsets by means of semantic

relations. However, a set of one or more specific relations for linking MWEs to meanings of the component words, as proposed by Osherson & Fellbaum (2010), has not been defined yet. On the other hand, the existence of aligned wordnets¹ for tens of languages offers easy access to MWEs in other languages and can serve as material for multi- and cross-lingual studies, as illustrated by Leseva et al.'s chapter.

Being concerned with the mapping of meaning to form via the theory of Frame Semantics (Fillmore 1976, 1977, 1982), the FrameNet lexical database (Baker et al. 1998) seeks to account for the semantics of lexical units by assigning them to semantic frames whereas the valences or combinatorial possibilities of each item are revealed from semantically and syntactically annotated sentences from which reliable information can be obtained. In this volume, Giouli et al. make use of FrameNet mechanisms for representing the semantics of MWEs in the light of their valences and the lexicon-corpus interface.

The development of MWE lexicons is intended both for automatic exploitation in NLP and for human usage. With respect to the former, the mere computational format of these resources shows that developers are aware of the need for automatic language processing, while a concern for standardization is proof of the language engineers' need to access such linguistic knowledge. However, tools for manual retrieval of MWEs from lexicons and even from corpora have been created and one of them is presented by Odijk et al. in this volume.

Hana Skoumalová, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondříčka, and Milena Hnátková present LEMUR, a MWE lexicon for Czech. The paper is an attempt to innovatively capture MWEs in Czech so that they can be annotated and searched for in large corpora, thus allowing the user to make effective use of them. Detailed properties concerning both the MWE as a whole and its components are included; for example, for MWEs, the types of idiomaticity (morphological, syntactic, semantic and statistical) are distinguished. At the same time, the entries are designed in such a way that the considerable variability of MWEs in the corpus texts (fragments, varied word order, syntactic modification, etc.) can be captured as well as possible, i.e. to include as many uses of variable MWEs as possible in the search. The MWEs annotated in the corpus are also linked to the corresponding entries in the database, where detailed searchable properties of the MWEs are available to the user, including

¹The word *wordnet* is used to refer to a “lexical knowledge base for a given language, modeled after the principles of Princeton WordNet” (see http://www.dblab.upatras.gr/balkanet/journal/20_BalkaNetGlossary.pdf). The form *Wordnet* is used for a particular such resource, e.g., the Bulgarian Wordnet or the Romanian Wordnet; the form *WordNet* is used only for the trademarked Princeton WordNet (see <https://wordnet.princeton.edu/>).

their meaning, traditional linguistic categorization, typical examples, etc. Linking the corpus to the database allows the user to work with the current language and, for example, to determine the frequency of occurrence of individual MWEs in the corpus. Linking this database further with other lexicographic resources is a natural next step.

Stella Markantonatou, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chiril, Dimitrios Karamatskos, Nikolaos Valeontis, and George Pavlidis present the challenges involved in collecting and representing MWEs for non-standardized language varieties, the focus being on Pomak, an endangered, non-standardized language variety of the East South Slavic dialect continuum. The chapter describes an openly available, online dataset of Pomak verbal MWEs, which were collected via fieldwork. The resource was developed with IDION, a web-based environment for the documentation of a wide range of syntactic, semantic, and stylistic properties of the expressions. Translations and usage examples of the Pomak expressions are provided along with a syntactic analysis in the Universal Dependencies framework. In the collected data both light verb constructions and idioms have been observed.

Svetlozara Leseva, Verginica Barbu Mititelu, Ivelina Stoyanova, and Mihaela Cristescu describe an empirically devised framework for the creation of linked bilingual computational lexicons of MWEs. The framework is applied to a bilingual (Bulgarian and Romanian) lexicon of verbal MWEs, which aims at providing a comprehensive description of their features in each of the languages under study. The MWEs, derived from the Bulgarian and the Romanian Wordnet, represent counterparts or translation equivalents of each other; while they are described according to the common principles and features adopted, the data in each language constitute a self-contained monolingual lexicon which may be developed independently. The description of each monolingual lexicon entry includes technical details necessary for cross-lingual linking and a rich linguistic description, on multiple levels. The work illustrates the applicability of a uniform description of MWEs to two languages from different families in a way that accounts for linguistic similarities and specificities. The resource can be enhanced to cover other levels and features of linguistic description, as well as expanded towards other languages.

Petya Osenova and Kiril Simov model MWEs in the framework of integrated lexical resources that would facilitate various NLP tasks. They use the notion of catena, an alternative to representing the structure of MWEs in lexicons, for the unified encoding of the grammatical, lexical and semantic information. This kind of approach is tree-oriented, thus providing better possibilities for handling

idiosyncrasies in comparison to the static methods. The tree representations follow the ideology of Universal Dependencies. MWE lexical entries have a layered structure, with a complexity modelled with respect to two important features of MWEs: discontinuity and fixedness.

One challenge while encoding MWEs for Natural Language Understanding applications is the representation of their semantics. Voula Giouli, Vera Pilitsidou, and Hephestion Christopoulos present a frame-based lexical resource for Modern Greek and the encoding of nominal and verbal MWEs in it. To better account for the deep semantics of these complex predicates, their argument structure (or valency) is identified and their lexical-semantic description is provided by means of assigning them to a frame and identifying their Frame Elements. Lexicon development is based on corpus evidence and the annotation performed. The authors discuss the difficulties encountered due to the nature of these complex predicates. They also discuss on the basis of discrepancies observed between single- and multiword lexical units assumed under the same frame in terms of Frame Elements assignment and syntactic realization.

Christian Chiarcos, Maxim Ionov, Elena-Simona Apostol, Katerina Gkirtzou, Besim Kabashi, Anas Fahad Khan, and Ciprian-Octavian Truică set out the challenges of modeling MWEs within linked data lexicons and demonstrate how OntoLex-Lemon, a de facto community standard for modelling and publishing lexical resources on the Semantic Web, can effectively address them. Their chapter can serve as a guide for users grappling with the complexities of MWE data modeling in linked data lexicons. The reader is presented diverse strategies for modeling MWEs via the different modules of OntoLex-Lemon, both individually and in combination. The aim is to match specific modeling strategies with particular use cases. This chapter not only presents recommendations, but also furnishes practical examples drawing from real-world use cases, at the same time featuring a comparative analysis of OntoLex and other pre-RDF vocabularies, exploring the advantages and disadvantages of the former for existing tools and potential downstream applications in modeling MWEs.

Jan Odijk, Martin Kroon, Sheean Spoel, Ben Bonfil, and Tijmen Baarda present MWE-Finder, an application that enables a user to search for MWEs in large Dutch text corpora. To cope with the discontinuity of MWE components, with their word order variation, the search engine takes into account the MWE grammatical configuration. Searches are made possible by using a canonical form, which is an implicit hypothesis on the properties of the MWE with regard to form variation, modification, and determination. To this end, the DUTch CANonicalised Multiword Expressions lexical resource (DUCAME) is used. The chapter presents an overview of DUCAME, demonstrates the user interface, describes

the redesign of the back-end needed for dealing with large text corpora, and illustrates the application for a specific MWE example showing how unexpected form variations, modifications, and determinations, as well as a variant of the MWE are found.

The development of computational models of compositionality typically goes hand in hand with the creation of reliable lexical resources as gold standards for formative intrinsic evaluation. Even though datasets of noun compounds with ratings on compositionality across languages have been developed for many languages, work that looks into whether and how much both the gold standards and the prediction models vary according to the properties of the targets within the lexical resources is still scarce. In her chapter, Sabine Schulte im Walde suggests a novel route to assess the interactions of compound and constituent properties concerning the degrees of compositionality of the compounds while focusing on English and German noun compounds. A novel collection of compositionality ratings for German noun compounds is proposed, where human judges were asked to provide compound and constituent properties before judging the compositionality. Also, a series of analyses on rating distributions and interactions with compound and constituent properties for the novel collection, as well as existing gold standard resources in English and German are made and discussed. The author recommends assessing computational models not only on the full dataset, but also on subsets of targets with coherent task-relevant properties.

Fluency in a (new) language comes from mastering the vocabulary and semantics, the rules for inflecting and combining words in phrases and sentences, the pragmatic factors, the cultural knowledge, but, to the same extent, from knowledge about the word combination possibilities (Ramisch 2023). Therese Lindström Tiedemann, David Alfter, Yousuf Ali Mohammed, Daniela Piipponen, Beatrice Silén, and Elena Volodina present part of a new resource, the Swedish L2 profile. It provides access to MWEs which can be filtered according to type and the level in the Common European Framework of Reference (CEFR) and includes receptive and productive statistics of usage in corpora, as well as links to the empirical data upon which the resource has been built. This makes the resource useful for research, teaching and technical developments. The experiments presented in the chapter show that the receptive difficulty of MWEs is evaluated similarly by experts and non-experts, while their level of compositionality or transparency influence their ranking on the CEFR scale.

After more than two decades since MWEs were initially discussed in the literature of Natural Language Processing (NLP), there are still open issues of all sorts, starting with the very definition of a MWE, as readers will also notice in the chapters of this volume. It was beyond our scope to have a common understanding

of this concept, as all phenomena covered are related to a certain extent and it is relevant to see how their descriptions can be leveraged with mutual benefits.




References

- Baker, Collin F., Charles J. Fillmore & John B. Lowe. 1998. The Berkeley FrameNet project. In *36th annual meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 1, 86–90. Montreal: Association for Computational Linguistics.
- Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod & Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the third international Conference on Language Resources and Evaluation (LREC'02)*, 1934–1940. Las Palmas, Canary Islands: European Language Resources Association (ELRA).
- Constant, Mathieu, Gülşen 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. DOI: 10.1162/COLI_a_00302.
- Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag & Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the third international Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Canary Islands: European Language Resources Association (ELRA).
- Evert, Stefan. 2004. *The statistics of word cooccurrences: Word pairs and collocations*. (Doctoral dissertation).
- Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database* (Language, Speech, and Communication). Cambridge, MA: MIT Press.
- Fillmore, Charles J. 1976. Frame Semantics and the nature of language. *Annals of the New York Academy of Sciences* 280. 20–32.
- Fillmore, Charles J. 1977. Scenes-and-frames semantics. In Antonio Zampolli (ed.), *Linguistic structures processing: Fundamental studies in computer science*, vol. 59 (Fundamental Studies in Computer Science), 55–81. Amsterdam; New York; Oxford: North Holland.
- Fillmore, Charles J. 1982. Frame Semantics. In *Linguistics in the morning calm: Selected Papers from SICOL-1981*, 111–137. Seoul, Korea: Hanshin Publishing Company.

- Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.
- Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.
- Osherson, Anne & Christiane Fellbaum. 2010. The representation of idioms in WordNet. In Pushpak Bhattacharyya, Christiane Fellbaum & Piek Vossen (eds.), *Principles, construction and application of multilingual wordnets: Proceedings of the fifth Global WordNet Conference*. Mumbai, India: Narosa Publishing House.
- Ramisch, Carlos. 2023. *Multiword expressions in computational linguistics: Down the rabbit hole and through the looking glass*. Aix Marseille Université (AMU). <https://theses.hal.science/tel-04216223>.
- Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the joint workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91. Florence. DOI: 10.18653/v1/W19-5110.

Chapter 1

LEMUR: A lexicon of Czech multiword expressions

 Hana Skoumalová^a,  Marie Kopřivová^a,  Vladimír Petkevič^a,  Tomáš Jelínek^a,  Alexandr Rosen^a,  Pavel Vondříčka^a &  Milena Hnátková^a

^aCharles University

This chapter describes a lexicon of Czech multiword expressions, designed to be useful both for human readers and for natural language processing tasks. Its entries use a rich typology of multiword expressions, based on their syntactic aspects, idiomatity and flexibility, with a focus on the specific features of Czech multiword expressions with their significant variability, and a classification according to a traditional approach. The content and structure of the entries facilitate the use of the lexicon in natural language processing. The chapter also describes how the lexicon is implemented and used in parsing and for annotating multiword expressions in a corpus. The corpus and the lexicon are linked, so each entry in the lexicon includes examples from the corpus, and each annotated multiword expression in the corpus is linked with a corresponding lexical entry.

1 Introduction

In every language, multiword expressions (henceforth MWEs) represent a substantial part of the vocabulary, both in common and in specialist use. A lexicographical resource describing MWEs is therefore an obvious need. Such descriptions can be part of a standard lexicon or included in a dedicated lexicon of MWEs.

On the path from lexicon to grammar, many MWEs stay at least halfway between the two, some much closer to grammar than single-word lexemes. This



is even more pronounced in a language such as Czech, with its free word order and rich morphology, including intricate morphosyntactic agreement patterns. Considering the flexibility of many MWEs (not only in Czech), allowing for insertions, omissions, permutations, morphosyntactic transformations, the use of synonyms, and other manifestations of variability, a satisfactory solution calls for a highly elaborate scheme for the specification of lexical entries. As an answer to the need for a lexical resource up to the task we introduce LEMUR, a LExicon of MUltiword expREssions of Czech.

The chapter is structured as follows. §2 relates LEMUR to some existing common lexical resources, referring to its sources of inspiration and providing a concise summary of research on MWEs, with a specific emphasis on Czech. The extensive §3 introduces the components of a lexical entry together with the multi-dimensional taxonomy of MWEs. Next, §4 presents an overview of how the lexical entries are encoded and how the whole lexicon is implemented. In §5 we exemplify the current use cases of the lexicon: (i) as a resource for annotating MWEs in corpora and providing links between their occurrences in a corpus and the corresponding entries in the lexicon, and (ii) as an aid in tagging and parsing. The chapter concludes with a summary of achievements and pitfalls and some perspectives of the project (§6).

2 LEMUR related to other MWE lexicons and previous research

LEMUR was designed from the start as a richly structured database, with an interface suitable for use in lexicography, for teaching Czech as a foreign language, for studying theoretical issues of MWEs as entities between lexicon and grammar, and also for Natural Language Processing (henceforth NLP) tasks such as tagging, parsing and corpus annotation, including MWE identification and search, or word sense and semantic disambiguation.

2.1 LEMUR and other MWE lexicons

LEMUR is not the first lexicon of Czech MWEs. The standard reference lexicon of Czech phraseology (Čermák et al. 1983–2009) is an impressively large and detailed achievement, but its printed format and standard lexicographical approach favour the traditional manual look-up before other possible uses. Other resources focus either on an inventory of MWEs used for their identification in corpora, such as the FRANTALEX lexicon (Hnátková 2002, Kopřivová & Hnátková 2014),

or on extending a valency lexicon to include MWEs headed by verbs (Urešová 2009, Lopatková et al. 2014, Przepiórkowski et al. 2017). LEMUR differs from the above in its broad focus: to the best of our knowledge, its entries cover more types of MWEs and capture more properties of each MWE than any other resource (for a similarly rich resource for Bulgarian and Romanian, see Leseva et al. 2024 [this volume]). Moreover, it provides the option of bi-directional links between entries in the lexicon and occurrences of the multiword lexemes in a corpus.

In addition to MWEs listed in traditional phraseological dictionaries, i.e. proverbs, similes and sayings (Burger et al. 2007), the lexicon includes compound function words (mainly prepositions and conjunctions), scientific terms (Kovářiková & Kovářik 2019), and typical collocates (for example *vydatná strava* ‘nutrient food’). However, it does not include frequent co-occurrences of function words such as *ale i* ‘but also’, *že se* ‘that REFL’ etc.

LEMUR builds on FRANTALEX, which was based on Čermák et al. (1983–2009) and extended by additional MWEs, found in corpora, and variants of already known MWEs. The MWE typology used in the lexicon is a modification of the multi-dimensional taxonomy used in lexical templates within the PARSEME project,^{1,2} and inspired by Baldwin & Kim (2010). An important addition to the taxonomy is the notion of morphological idiomaticity (see §3.3.2). While compiling the lexicon, we also addressed theoretical issues related to the variability of MWEs (Pasquer et al. 2018). As a major theoretical contribution, we see the design of a scheme describing the variability, together with detailed descriptions of variability of each MWE.

Last but not least, the entries include the syntactic structure of each multiword lexeme as dependency and constituency trees. This view of MWEs is important also for the section of the entry where possible valency requirements of a part or the whole of the MWE may be specified.

2.2 MWE research mainly from the Czech perspective

Research on MWEs intensified in the late 20th century. Properties and usage of MWEs have been studied from various angles. Some of the studies deal mainly with terminology (Bozděchová 2007, Temmerman 2000), while non-compositional MWEs are studied within the disciplines of phraseology, paremiology (Čermák 2007), and also comparative studies (Popovičová 2020: 12–16). With the de-

¹<https://typo.uni-konstanz.de/parseme/>

²https://www.lexical-resource-semantics.de/wiki/index.php/Parseme_MWE_Template:_English (visited Nov 13, 2023)

velopment of language corpora, new possibilities for research on MWEs and collocations emerged: for terms (Kovářiková 2017), and for phrasemes (for example Colson 2017). A new concept of terminology (Klégr 2016) and new demands for the identification of MWEs within large-scale data appear. In recent years, NLP focusing on the description of MWEs has become one of the fastest growing areas, obviously relevant also for Czech (Lichte et al. 2019, Sheinflux et al. 2019).

The description of phraseology is essential for language teaching (Čechová 2011: 66–67), lexicography (Čermák et al. 1983–2009) and linguistic theory. In order to handle concrete data, the properties of phrasemes become important. However, individual researchers differ even here (for example Čechová 2011, Čermák et al. 1983–2009): for Čechová, a MWE is characterized by the fixedness of form, while Čermák allows for the possibility of variability in some MWEs. We see variability as a complex phenomenon: some MWE properties, such as variation, fixedness and repetition of a lexeme, need to be described in a way more consistent with real-text data (Jelínek et al. 2018). Thanks to the availability of large corpora, variation and fixedness can be observed in more detail to see where grammatical categories alternate and where new MWEs with new lexical components emerge: for example plural and singular alternate in *cesta do pekla/pekel*, (lit. ‘way to hell/hells’), while a new MWE *dávat logiku*, (lit. ‘to give logic’), is derived from its original version *dávat smysl*, (lit. ‘to give sense’). Corpora also help to identify and annotate monocollocable components within MWEs, i.e. words with restricted usage in one or few combinations only, and to mark MWEs containing such components. Moreover, in our approach we also annotate fragments of MWEs since MWEs often occur in fragmentary forms.

In the LEMUR lexicon, we also try to reconcile the different approaches and introduce a taxonomy of MWEs that encompass different linguistic domains. This makes it possible to search for units according to criteria used by different approaches, with an emphasis on the traditional Czech MWE categorization that reflects the current educational needs. In order to classify MWEs, we use a combination of the classification presented by Čermák (2007) and Moon (2007), with the addition of some new categories. In determining the type of idiomaticity, we adopt the PARSEME taxonomy, supplemented with categories related mainly to morphology (Hnátková et al. 2017).

3 Typology of MWEs

The MWE typology in LEMUR is inspired by the PARSEME project and by Baldwin & Kim (2010), which categorizes MWEs according to their

- *idiomaticity*: lexical, syntactic, semantic, pragmatic and statistical;
- *syntactic category*;
- *fixedness/flexibility* of “lexicalized phrases”.

This typology was adopted and primarily extended with respect to: (i) specific properties of Czech, especially morphological idiomaticity, and (ii) the fact that the lexicon should be useful both for human users and software applications. For instance, in (1) the form *nosa* ‘nose’ is a nonstandard genitive form of the noun *nos* appearing exclusively in this MWE (cf. also §3.3.2). This fact is marked on *nosa* in the lexical entry.

- (1) podle **nosa* poznáš kosa
 by nose.SG.GEN recognize.2SG.PRS blackbird.SG.ACC
 ‘someone’s character can be recognized by her/his deeds’

Moreover, we also extended PARSEME’s approach in the following respects. In our approach, lexical idiomaticity (§3.3.1) encompasses not only MWE components that are not part of the conventional lexicon of Czech, such as MWEs consisting of foreign loans, for example (LAT) *mutatis mutandis*, but also (possibly almost) monocollapsible words, for example *překot* in *o překot* ‘headlong’, negative only forms, for example *neličená radost* ‘genuine pleasure’, macaronic structures, that is, structures combining Czech and foreign words, for example *by voko* ‘by guesswork’ and other. Syntactic idiomaticity is not restricted only to MWEs whose syntax is not derived from that of their components, since we also annotate their deviations from standard syntax, such as anacolutha (cf. 31), attraction (cf. 32), idiosyncratic valency (cf. 33), aposiopesis (cf. 34), ellipses (cf. 35), zeugmas and others. For capturing semantic idiomaticity, we use a 4-grade annotation scale where a MWE can be: (i) always non-compositional, i.e. not explicitly derivable from its components, for example *nebrat si servítky* (lit. ‘not to take napkins’), ‘not to mince one’s words’), (ii) rarely compositional, for example *kukaččí vejce* ‘cuckoo’s egg’, (iii) often compositional, i.e. often non-idiomatic, for example *vlčí doupe* ‘wolf’s den’, and (iv) always non-idiomatic, literal, for example *přísný pohled* ‘stern look’. As for syntactic structure (§3.1), MWEs are identified by their syntactic category (determined by MWE’s head) and assigned a dependency and a constituency tree. Moreover, the (im)possibility of MWEs’ syntactic transformations (passivization, nominalization, adjectivization) is also annotated. As to MWE fixedness and flexibility, MWEs are specified for the (im)possibility

of their lexical, morphological/morphosyntactic and syntactic variation, including internal modification of their components and/or word order fixedness or freeness.

Generally, each lexicon entry contains descriptions of two types of MWE properties: some concern the MWE as a whole, others are related to its components (words). In addition to the three characteristics adopted from the PARSEME project, each entry in the lexicon is described by its lemma and superlemma, definition, examples of usage, style marker, usage type, i.e. a collocation type classified according to the traditional Czech phraseological taxonomy, and a basic classification of adverbial MWEs. The detailed description of these features follows.

- Lemma is a string (= sequence of concatenated word forms) constituting a MWE in its prototypical form that identifies the MWE in the lexicon and in an annotated corpus, for example *materí kašička* ‘royal jelly’. Via its lemma, the MWE can be searched both in the lexicon and in the corpus.

In the case of synonymous MWEs, we have to decide whether they will be included under one lemma or whether the lexicon will contain two lemmas. If they differ only in meaning, but all other properties are the same, for example, *černá díra* ‘black hole’ (a scientific term vs. collocation meaning that money disappears somewhere with no visible benefit), the dictionary will contain only one lemma with two definitions and two sets of examples. However, if there is variability in the lexical setting of one of the lemmas, or constraints on syntactic transformations, word order changes, etc., we will introduce two lemmas, as in (2).

- (2) *jít přes čáru*
go over line
‘to cross the border illegally/to cross the line’

In the meaning ‘to cross the border (illegally)’, we can use synonyms of the verb *jít* ‘go’ that differ in aspect or prefix: *jít/přejít/přecházet/chodit přes čáru*. In the meaning ‘to behave in an unacceptable way’, only the imperfective aspect of the verb *jít* is possible, but the verb *být* ‘be’ can also be used here to describe the state when someone has crossed an imaginary boundary of decency (*jít/být přes čáru*).

- Superlemma is a representative of a list of lemmas that have at least one word in common and are semantically related. These are, for example, converse MWEs such as (3a) and (3b), or two related, but different MWEs such

as (4a) and (4b). Note that (4b) is not a standard nominalization of (4a) (unlike (4b), (4a) is a comparison/simile containing the conjunction *jako* ‘like’),³ but both share the same superlemma.

- (3) a. *dát ultimátum*
 give ultimatum
 ‘to give an ultimatum’
 b. *dostat ultimátum*
 get ultimatum
 ‘to get an ultimatum’
- (4) a. *hrát si jako kočka s myší*
 play REFL like cat with mouse
 ‘to play like a cat with a mouse’
 b. *hra kočky s myší*
 play of cat with mouse
 ‘a cat and mouse game’

A superlemma is always an existing lemma, or a fragment thereof and must consist of at least two words. For the examples given, the superlemmas are *dát ultimátum* ‘give an ultimatum’ and *hra kočky s myší* ‘play of a cat with a mouse’, respectively. The superlemma is actually only a label indicating a list of semantically linked lemmas, and we select the shortest lemma or fragment from the list as the superlemma.

- Definition is an informal gloss of the MWE’s meaning. Most glosses are adopted from Čermák et al. (1983–2009).
- Examples are from corpora of contemporary Czech, representing real usage.
- Stylistic marker classifies both the MWE and its components (words) from the viewpoint of style. The following values are distinguished:
 - standard: used commonly in written texts: *být upoután na lůžko* ‘to be confined to bed’;

³The standard nominalization would be *hraní si na kočku a myš* where *hraní* ‘playing’ is a paradigmatically derived deverbal noun.

- colloquial: used mainly in spoken communication and understandable in every part of the country: *Co jí to vlezlo do hlavy?* (lit. ‘What crept into her head?’) ‘Where did she get that idea?’;
- dialect: the whole MWE or one of its components is part of a particular dialect spoken only in part of the country. Such a MWE is included in the lexicon since it occurs in corpus texts (in fiction or regional newspapers); dialect phraseology as such is not included in the lexicon. For instance, in (5) *náčeňovou hadrou* ‘(with a) dish cloth’ is a dialectal expression;
- slang: *naprat to pod klacek* (lit. ‘to nail it under the stick’) ‘to hoof the ball into the net’;
- other: literary expressions, mainly from the Bible or classical (Greek, Roman) literature, and other sayings: *překročit Rubikon* ‘cross the Rubicon’.

- (5) lepší než náčeňovou hadrou přes papulu
better than dish.INS cloth.INS over gob
‘better than a poke in the eye with the sharp stick’

In addition to the above categories, every word and every MWE can be marked as having an expressive meaning. The words *rachot*, *bengál*, *varvas*, *bordel* denote ‘rumble’ in different styles and all are expressive. On the other hand, *vylít někomu boty* (lit. ‘pour out one’s shoes’) ‘throw someone out on their ear’, consists of non-expressive terms, but the entire MWE is expressive.

Generally, the style values are mutually exclusive except for the expressive value that can be assigned to a word or to a MWE together with some other value.

- Usage type is based on a classification common in the Czech linguistic literature (Čermák 2016) and the lexeme-specific data from Čermák et al. (1983–2009). The following values are distinguished:
 - proverb: *Chybovat je lidské*. ‘To err is human.’;
 - weather lore: a traditional saying used to predict or interpret weather patterns, or to suggest what people should do on certain dates (6);

- comparison/simile: a collocation typically formed by a verbal or adjectival phrase containing an expression to which something is compared (7);
 - citation: part of another text presented verbatim and taken over from literature, film etc.: *Knihy mají své osudy*. (lit. ‘Books have own fates.’) ‘Books have their own destiny.’;
 - foreign collocation: a collocation taken over unchanged from a foreign language (typically Latin, Greek, English, German, French): (FRE) *raison d’être*, (ENG) *by the way*;
 - scientific/professional term: *diferenciální rovnice* ‘differential equation’;
 - multiword function words: used mostly as prepositions or conjunctions: *bez ohledu na* (lit. ‘without regard to’) ‘regardless of’;
 - (non-specific) verbal MWE: a semantically non-compositional MWE including a verb form as its governor (8);
 - non-verbal MWE: a semantically non-compositional MWE not including a verb: *něžné pohlaví* (lit. ‘gentle sex’) ‘the fair sex’;
 - quasiphraseme: collocation composed of an abstract noun and one of the very limited set of phase verbs (inchoative, durative, terminative): *věnovat pozornost* (lit. ‘donate attention’) ‘pay attention’; it is usually difficult to find single-verb equivalents for these MWEs;
 - sentential phraseme: a phraseme differing from a proverb, a weather lore or a citation: *a co ty?* (lit. ‘and what you?’) ‘and what about you?’;
 - open phraseme/set phrase: a MWE requiring a continuation, typically routine formulation introducing a text or conversation (Coulmas 1981, Aijmer 1996), which is typically further expanded: *jen si představte...* ‘just imagine...’;
 - (usual) collocation: a collocation based on semantic/selectional restrictions only: *úhlavní nepřítel* (lit. ‘principal enemy’) ‘arch-enemy’;
- (6) Na svatého Jiří vylézají hadi a
 On saint.GEN George.GEN creep out snakes.NOM and
 štíři.
 scorpions.NOM.
 ‘On Saint George’s Day the serpents and scorpions creep out.’

- (7) líný jako veš
lazy like louse
'lazy as a bear'
- (8) na tom nesejde
on it descend.NEG.3SG.PRS
'it makes no difference'

- Adverbial MWEs are classified by the four basic semantic categories (place, time, manner, circumstance):
 - adverbials of place: *na pokraji* 'on the brink of';
 - adverbials of time: *dnem i nocí* 'day and night';
 - adverbials of manner: *po vzoru* 'on the pattern of';
 - adverbials of circumstance: *u příležitosti* 'on the occasion of'.

3.1 Syntactic structure

As another feature inspired by the PARSEME project, each entry is characterized by its syntactic type, i.e. a syntactic category it constitutes in the sentence: NP, AdjP, VP (distinguishing content verb and categorial/light verb phrases), AdvP, PP, or compound preposition/conjunction/interjection, clause, compound sentence. Moreover, for every MWE, the lexicon specifies its dependency and constituency structure, both represented as syntactic trees. Another possible way to capture the syntactic structure of the MWE is a catena (Osenova & Simov 2024 [this volume]). Dependency trees (including syntactic functions) are produced by a parser. In the past, it was TurboParser (Martins et al. 2013), but today it is a parser from the NeuroNLP2 tools (Ma et al. 2018). The parses are manually checked, and then constituency trees are derived from dependency trees using a rule-based conversion system.

Whenever a MWE requires some of its parts to be a lexically unspecified constituent, the syntactic head (verb or adjective) is provided with information on its valency (Rosen & Skoumalová 2018). If necessary, entries may specify the valency of the whole MWE. This is the case, for example, for some constructions consisting of a verb and a nominal or prepositional object: they may take a complement which is required neither by the verb nor by the object. Thus, the MWE in (9) can be complemented, for example, by a *that*-clause, while such a clause can complement neither the verb *dát* 'give' nor the noun *srozuměnou*.

- (9) dát na srozuměnou
give on understanding.SG.ACC
'to let know'

Thus, each MWE is described by its syntactic structure, syntactic type and – if syntactically non-standard – also by the kind of its syntactic idiomaticity (see §3.3.3).

3.2 Variability/flexibility

Variability is understood in several different meanings (Hnátková et al. 2017):

- lexical variability: some positions in a MWE can be occupied by synonyms;
- morphological variability: a MWE can possibly occur in various morphological forms;
- word order variability: specific/anomalous free or fixed word order within (parts of) a MWE;
- syntactic transformations: passivization, nominalization, adjectivization, etc.;
- insertion of modifying elements in between the standard MWE template/pattern, i.e. syntactic *modifiability* of MWE components;
- omission of words resulting in *fragments* of standard MWEs.

Unless specified otherwise, we assume that MWEs behave in the same way as regular constructions and contain morphologically standard forms. Hence, we only indicate violations of default properties and rules of grammar. For instance, one of the general properties in Czech is its free word order, thus only specific word order configurations in MWEs are indicated in their lexicon entries.

It is important to account for variability on various levels of linguistic description since one of the objectives of the lexicon is to make it possible to identify not only MWEs in their standard, canonical forms (expressed, for example, in their lemmas) but also their modifications of various kinds. It is often the case that language users modify standard MWEs in a creative way. The lexicon entries cover the kinds of variability listed in §3.2.1, §3.2.2, and §3.2.3 below.

3.2.1 Lexical variability

Lexical variability can be indicated in each lexical position, where appropriate, ranging from a specific word to a choice of several variants (for example synonyms) to a completely free choice determined only by an appropriate word class. A special case of this type of variability is a MWE where a certain lexeme is repeated, while this lexeme can be chosen from several variants (Jelínek 2020). For instance, in the Biblical saying (10) the lexeme *Bůh* ‘God’ is repeated in the second clause, which has the opposite meaning. This MWE can be seen as a template where both positions occupied by *Bůh* ‘God’ are in fact containers that might be filled with (almost) arbitrary, but identical nouns (for example *Život dal, život vzal* ‘Life gave, life has taken away’; *Bolševik dal, bolševik vzal* ‘Bolsheviks gave, Bolsheviks has taken away’).

(10) Bůh dal, Bůh vzal.

God gave, God took.

‘The Lord gave, and the Lord has taken away.’ (the Book of Job 1,21)

Single-word lexical synonyms within a MWE can sometimes have the form of a multiword microstructure, for example a non-reflexive verb can be expressed by its reflexive synonym consisting of a verb and its reflexive particle (*se/si*) as a free morpheme, which need not occupy an adjacent position. This results in different syntactic structures of MWE’s synonymous variants and may complicate a successful identification of such a MWE in texts.

3.2.2 Morphological variability

Due to the rich morphological system of Czech, MWEs can occur in various morphological forms, for example verbs can differ in aspect (perfective, imperfective, biaspectual), nouns can appear in various cases or numbers, adjectives or adverbs can occur in the comparative or superlative degree, etc. The morphological richness is illustrated by examples (11), (12) and (13). The following MWE represented by the same lexical entry can appear in two variants, reflected in the entry: in the nominative plural *houby* ‘mushrooms’ or genitive plural *hub*:

(11) přibývat jako houby/hub po dešti
multiply like mushroom.PL.NOM/GEN after rain.
‘to spring up like mushrooms’

For instance, the MWE *nebrat konce* ‘to be no end to [something]’ can appear in two variants: the noun *konec* ‘end’ is typically in the genitive of negation

(*konce*), but it can also, rarely, be in the accusative case (*konec*), satisfying the object valency requirement of the transitive verb *brát* ‘take’:

- (12) pořád to *nebere* *konce/konec*
 always it take.NEG.3SG.PRS end.SG.GEN/ACC
 ‘there is no end to it’

Paradoxically, fossilized constructions with the obsolete genitive of negation are typical examples of morphological variability.

Verbs in Czech, as in other Slavic languages, express aspect lexically. For instance, a single lexical entry can include both the perfective and imperfective variant of a verb:

- (13) *koupit/kupovat* něco *za babku*
 buy.PFV/IPFV something for old woman
 ‘to buy something dirt cheap’

Since aspect is a lexical rather than morphological category, aspectual variability is treated as lexical variability. For instance, there are MWEs permitting only one aspectual variant of a verbal lexeme: in (8) the perfective verb form *nesejde* ‘descend’ cannot be replaced by its imperfective counterpart *neschází*.

3.2.3 Word order variability

Although free word order is a typical trait of Czech, constructions with a fixed word order do exist: the position of prepositions within prepositional phrases, the position of prepositional phrases within noun phrases or the position of clitics within clauses or sentences. Free word order applies to clausal constituents. In the entries, only anomalies concerning both free and fixed word order are captured. For instance, in a MWE consisting of a verb and its syntactic object (14) the verb *dělat* and its object noun *aféru* can appear in either order – this regular syntactic fact is not recorded in the lexicon entry.

- (14) *dělat* z něčeho *aféru*
 make from something affair.ACC
 ‘to make a big deal about something’

Word order variations can also be due to standard grammar rules (concerning, for example, the position of clitics) or topic-focus articulation. The MWE in (15a) appears in sentence (15b) where the verb and the reflexive particle, components

of the inherently reflexive verb *rovnat se* ‘match’, occur in the reversed order, separated by the verb form *nemohla* ‘could not’. Again, this standard grammatical word order is not indicated in the entry.

- (15) a. *nemoci se rovnat*
can.NEG.INF REFL match
‘to be no match for somebody’
- b. *Co se týče rozpočtů, se nepálská studia nemohla indickým rovnat.*
What REFL concern.3SG.PRS budgets, REFL Nepali studies
can.NEG.PST Indian.DAT match.INF
‘In terms of budgets, Nepali studies could not match Indian studies.’

On the other hand, in the anomalous syntactic structure in (16) the noun *slova* ‘word’, an attribute in the genitive case, precedes its syntactic head *smyslu* ‘sense’; this kind of reversed word order is very rare and appears – as well as other word order anomalies – primarily in MWEs, duly indicated in their lexical entries.

- (16) *v nějakém slova smyslu*
in some.LOC word.GEN sense.LOC
‘in some sense of the word’

3.2.4 Syntactic transformations

MWEs related by the same or similar meaning can appear in various syntactic structures that are derived by syntactic transformations from a basic variant. We account for transformations of the following three types, marking only the structures and patterns that are idiosyncratic with respect to the standard grammar of Czech:

- Passivization/depassivization. The following features can be specified in lexical entries:
 - MWE cannot be passivized: a flag specified for MWEs headed by a transitive verb that cannot be passivized in this particular MWE (as an exception to the general rule stating that every transitive verb can be passivized). For instance, the verb *spatřit* ‘see’ can be passivized in general, but cannot be passivized in (17).
 - MWE cannot occur in the active form, for example the MWE in (18) exists only with the passive form *přáno* ‘wished’.

- (17) spatřit světlo světa
 see light world.GEN
 ‘to come into the world’
- (18) nebylo mu přáno
 be.NEG.3SG.N.PST he.DAT wish.PASSP
 ‘he was out of luck’

- Nominalization. We assume that a verb in a MWE can be nominalized. If this is not the case, such a verb is flagged appropriately. For instance, in (19) the reflexive verb *hodit se* ‘be suitable’ cannot be nominalized and this negative fact is recorded in the MWE.

- (19) hodit se jako pěst na oko
 be suitable REFL like fist on eye
 ‘to be completely out of place’

- Adjectivization. Similarly as with nominalization, it is assumed that generally a verb in a verbal MWE can be adjectivized. If not, such a MWE is marked appropriately. In (20), the impersonal neuter verbal participle *došlo* ‘it got to’ cannot be adjectivized and this fact is duly recorded as this MWE’s property.

- (20) *došlo na má slova*
 get.3SG.N.PST on my words
 ‘my words came true’

3.2.5 Insertion

Normally, content words within MWEs can be syntactically modified; typically, adjectives modify nouns, adverbs modify verbs, adjectives or adverbs, etc. Such regular syntactic structures are not reflected in the annotation of MWEs. For instance, the MWE in (21a) can appear in a text as in (21b).

- (21) a. *nechávat si něco pro sebe*
 keep REFL.DAT something.ACC for oneself
 ‘to keep something to oneself’

- b. Navrátil *si* *krutou* *informaci* *nechával* *dlouho* *jen*
Navrátil REFL.DAT cruel.ACC information.ACC keep.PST long.ADV only
pro sebe.

for himself.

‘Navrátil kept the harsh information only to himself for a long time.’

In (21b), the modifying adverbs *dlouho* ‘for a long time’ and *jen* ‘only’ are inserted in between the components of the standard MWE.

However, there are MWEs whose components cannot be modified, i.e., no insertions in between their components are allowed: this fact is specified in MWE entries where appropriate. For example, in the MWE *něco k snědku* ‘something to eat’ the monocollocable noun form *snědku* cannot be modified. There are words, however, such as *příslivečný* ‘proverbial’ or *doslova* ‘literally’, which can modify almost any MWE of the appropriate syntactic category. Indeed, lexical entries do not specify the availability of such insertions.

3.2.6 Omission/Fragments

MWEs can sometimes appear in their reduced forms – as *fragments*, with the same meaning as the entire MWEs. Our ambition is to recognize MWEs not only in their full, canonical form but also in their partial, fragmentary form. For instance, the lexicon contains the following entry in its standard form:

- (22) *hoří* *někomu* *koudel* *u zadku*
burn.3SG.PRS somebody.DAT oakum.NOM at backside
‘somebody is in a tight corner’

Such an entry contains information on possible fragments (represented by identifiers of individual words and of (sub)structures) as central, nuclear parts of the MWE. This approach enables the user to identify even fragmentary MWEs in a text. Thus we can find fragments of standard MWEs such as (23), where only the fragment *hoří koudel* ‘burns oakum’ remains while the sequence *u zadku* ‘at backside’ is missing.

- (23) *Měl* *asi* *pocit, že mu* *hoří* *koudel*
have.3SG.PST probably feeling, that he.DAT burn.3SG.PRS oakum
kvůli *Karlovi.*
because of Karel.

‘He had a feeling that he is in a tight corner because of Karel.’

In some MWEs there may be two representative fragments that allow for identifying such MWEs in texts. For instance, the standard MWE (24) contains two fragments that might identify the original full-fledged standard MWE: (i) *mazat [někomu] med* ‘spread [someone] with honey’, (ii) *med kolem huby* ‘honey around the gob’. Both fragments are marked in the entries of such MWEs.

- (24) *mazat někomu med kolem huby*
 spread somebody.DAT honey around gob
 ‘soft-soap someone/butter someone up’

In this way, we also capture various modifications leading to reduced versions of standard MWEs, reflecting the authors’ creativity.

3.3 Idiomaticity

We stick to the definition of idiomaticity proposed by Baldwin & Kim (2010), adopted also in the PARSEME project:

In the context of MWEs, idiomaticity refers to markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels. A given MWE is often idiomatic at multiple levels... (Baldwin & Kim 2010: 4)

In particular, we distinguish between lexical, morphological, syntactic, semantic, pragmatic and statistical idiomaticity. The types of idiomaticity used in the PARSEME project were extended by morphological idiomaticity to capture Czech word forms which do not exist outside the specific MWEs. For instance, in the MWE (25) the adjective *pitomá* ‘stupid’ is a non-inflected feminine form, but in the MWE it is used as an expressive form that morphologically does not agree with the masculine noun form *kluk* ‘boy’:

- (25) *kluk pitomá*
 boy.NOUN.M stupid.ADJ.F
 ‘stupid boy’

Below, the types of idiomaticity are described in detail.

3.3.1 Lexical idiomaticity

Lexical idiomaticity concerns MWEs containing lexically idiomatic word forms or lexemes. The following kinds of lexical idiomaticity are distinguished:

- Monocollocable word forms (26). The word *zadost* ‘satisfaction’ can exist in this MWE only. Such monocollocable words are often components of terms such as *kysličník osmičelý* ‘osmium tetroxide’.

(26) učinit zadost
make satisfaction
‘do justice’

- Almost monocollocable word forms, i.e. forms associated with a very limited set of collocates: *zorný úhel* ‘angle of vision / point of view’.
- Negative only word forms (27).

(27) nedílná součást
undivided part
‘integral part’

- Foreign loans: for example (28), a collocation loaned from German, phonetically and orthographically modified.

(28) mírnyx týrnyx
mir nichts dir nichts (GER)
lit. ‘nothing to me, nothing to you’
‘casually / as if it was nothing’

- Macaronic structures: for example, the following collocation consisting of the Latin preposition *per* and the Czech noun *huba* assigned the Latin morph *-m* (29).

(29) per *huba-m
via.LAT gob.CZE.F-LAT.F.SG.ACC
‘orally / by word of mouth’

- Other, such as verbatim translations: *potřást hlavou* ‘shake one’s head’ instead of *zavrtět hlavou* ‘turn one’s head’, or adaptations of foreign loans: *mandatorní výdaje* ‘mandatory expenses’ instead of *závazné/povinné výdaje*.

In the lexicon entry, every lexically idiomatic word form in a MWE is marked. Moreover, a single idiomatic form can be marked with multiple kinds of lexical idiomaticity at the same time. In the lexicon, we also plan to mark each MWE as containing/not containing a lexically idiomatic word form.

3.3.2 Morphological idiomaticity

Morphological idiomaticity concerns a morphologically non-standard morphological form existing only within a MWE. For instance, in the MWE *chca nechca* ‘nolens volens’, the forms *chca*, *nechca* are non-standard, the standard forms being *chtě nechtě* with the same meaning.

Similarly to lexical idiomaticity, every morphologically idiomatic word form in a MWE is indicated. We also plan to mark each MWE as containing/not containing a morphologically idiomatic word form.

Forms used in MWEs are sometimes licensed by rhyme, as in (30), where *sloupích* ‘columns’ is a non-standard variant of the standard form *sloupech*.

- (30) jména hloupých na všech *sloupích
 names stupid.PL.GEN on all.PL.LOC columns.PL.LOC (intended)
 ‘names of the stupid are on all columns’

3.3.3 Syntactic idiomaticity

Syntactic idiomaticity accounts for the following kinds of syntactic anomalies always concerning the entire MWE. They are marked on the MWE where appropriate.

- Anacoluthon: as in the modified New Testament saying (31).

- (31) Kdo po tobě kamenem, ty po něm chlebem.
 who at you stone.INS, you at him bread.INS.
 lit. ‘Whoever throws a stone at you, offer him bread.’
 ‘Do not repay anyone evil for evil.’

- Attraction: as in (32), where the imperative form *padni* ‘fall’ is repeated in the subordinate clause *komu padni* ‘to-whom fall’. The entire construction follows the *imperative wh-word imperative* template, which is realized by several different phrasemes.

- (32) Padni komu padni.
 Fall.IMP who.DAT fall.IMP.
 ‘Come what may.’

- Idiosyncratic valency: for instance, a noun in the obsolete genitive of negation as object of a negated transitive verb, the standard form being in the accusative case (33).

(33) nemám námitek
have.NEG.1SG.PRS objections.GEN
'I have no objections'

- Aposiopesis: unfinished sentence, as in (34).

(34) Já bych tě nejradši ...
I would you.ACC most preferably
'As for you, I wish I could...'

- Ellipsis:⁴

(35) Nevím, co [mám dělat] dřív.
Know.NEG.1SG.PRS what [have.1SG.PRS do.INF] sooner.
'I do not know what to do first.'

- Idiosyncratic word order: for instance, an adjective exceptionally (with respect to the grammatical system of Czech) follows its nominal syntactic governor: *mše svatá* (lit. 'mass holy').
- Other: ungrammatical/non-standard syntactic structures, contaminations, zeugmas, etc., such as (36), where the verb form *nevidím* 'I do not see' immediately follows a preposition *od* 'from' and *do* 'to', respectively, thus forming an ungrammatical structure:

(36) od nevidím do nevidím
from see.NEG.1SG.PRS to see.NEG.1SG.PRS
lit. 'from I can't see till I can't see' | 'all the time / without interruption'

3.3.4 Semantic idiomaticity

Semantic idiomaticity concerns a MWE's semantic (non-)compositionality, i.e. (non-)metaphoricity, viewed as the relative frequency of how often the MWE also appears in its compositional/literal meaning (as to the degree of compositionality of nominal MWEs, cf. also Schulte im Walde (2024 [this volume])). We use the following scale:

⁴The brackets in example (35) are used for marking the ellipsis.

- MWE is always non-compositional, i.e. always idiomatic – the situation described by the MWE can never happen in the real world:

(37) mít ocelové nervy
'to have nerves of steel'

- MWE is rarely compositional, i.e. it is often idiomatic:

(38) *strouhat* někomu *mrkvičku*
grate somebody.DAT carrot
'to express Schadenfreude'

- MWE is often compositional, i.e. rarely idiomatic:

(39) hrát si na schovávanou
play REFL at hide and seek
'to play hide and seek'

- MWE is always compositional, i.e. non-idiomatic, literal:

(40) dlouhodobá investice
'long-term investment'

3.3.5 Pragmatic idiomaticity

A MWE is pragmatically idiomatic if it is used in specific situations. For instance, a standard invitation to a dance sounds as in (41).

(41) Smím prosit?
May.1SG ask?
'May I have the pleasure (of this dance)?'

3.3.6 Statistical idiomaticity

Usual, frequent, semantically non-idiomatic collocations reflecting selectional restrictions in usage fall within this category. Some of their components have a very limited collocability potential. The components of such MWEs can hardly be replaced by synonyms, for example *vydatný déšť* 'heavy rain', or similarly in (42) where the adjective *dezolátní* is unlikely to be replaced by a synonym. In addition to usual collocations, we regard as statistically idiomatic also terms such as *bezkontextová gramatika* 'context-free grammar', and multiword function words (multiword prepositions and conjunctions).

- (42) být v dezolátním stavu
 be in desolate state
 'be in a state of neglect'

4 Design of the database

4.1 Basic data model

For full flexibility required by the potential variability of the expressions (see §3.2), we define the entry pattern by means of *slots* and *fillers*.

The entry unit consists of *slots* and *features* referring to the MWE as a whole. Slots represent the components of the MWE (pattern), which is the syntagmatic dimension of the MWE. Slots consist of *fillers* and the slot-specific *features*. Fillers represent the paradigmatic dimension of the components: the possible variants which may be used to realize a particular component (slot). The primary role of fillers is to represent actual (terminal) tokens to be matched in the data. They are defined by means of a combination of token attributes and their values that must be matched in the text data in order to identify the MWE as a whole (for slots and fillers see examples in §4.3.2). Other possible restrictions, such as those concerning word order, modifications or transformations, can be defined by means of additional features. Figure 1 shows the scheme of the entry structure.⁵

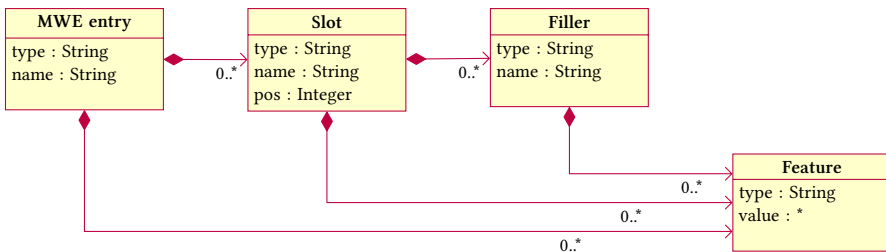


Figure 1: MWE entry structure (basic model)

All the container objects (entries, slots and fillers) have an arbitrary *name* and a *type*. Types are defined as a path in a hierarchy of categories defined in a separate metadata database. This helps to achieve a better organization and systematization of object types. For example, the atomic features may be easily classified by

⁵The structure follows (in a simplified form) basic principles of the proposal for a structured lexical description presented first in Vondříčka (2014) and has been described in full detail in Vondříčka (2019). In the current version of the database, the structure has been further simplified mainly by replacing filler attributes and references by dedicated features.

the linguistic layers they belong to (form, morphology, syntax, semantics, pragmatics, statistical properties, etc.). At a different level of classification, they can be grouped, for example by a particular purpose, linguistic theory or relative to a particular corpus. This also allows us to store multiple similar features from different sources or for different purposes at the same time. Features can easily be used (and classified) both for purely technical purposes of NLP processing tools and for storing information aimed at human users of the database, such as definitions, examples or notes.

In case we need to include multiple alternative values of some type of feature, additional custom subspecification may be used. This applies especially to user notes, examples from real texts or statistical values. For example, the basic type of feature for absolute frequency `:stats:freq:abs` is expected to be extended by additional custom subspecification of the corpus (and possibly subcorpus) used to acquire the frequency value, for example `:stats:freq:abs:BNC:fiction`. This allows the database to be searchable by features using underspecification of the type (by means of a path prefix) or its full (sub)specification as needed.

As described above, the fillers are expected to match more or less specific tokens in the text data. In our case, the data is already morphologically analyzed and disambiguated. This allows for underspecification of token attributes to be matched by using incomplete matching patterns or even regular expressions. We can, for example, match some lemma generally, independently of its particular morphological form (which is especially useful for verbs), or we can restrict the form more finely to some specific morphological category (number, case, person, etc.). It is also possible to match just some particular part of speech, such as adjective, demonstrative pronoun, etc. In case of specific valency requirements, it would be even more practical to match just whole syntactic phrases of a particular type instead of listing all their possible morphological realizations. While this is also possible in principle, unfortunately, we do not have a syntactic parser for Czech reliable enough to build upon. Therefore we need to identify MWEs solely by their realization in form of tokens on the surface level.

4.2 More advanced variability and structures

The basic model as described above makes it possible to deal with variability only at the level of single tokens, since one slot only corresponds to a single token possibly realized by various (single token) fillers. However, variability often concerns multiple tokens: prepositional phrases, periphrastic word forms, etc.

Instead of implementing a recursive database structure, we decided to keep it flat for practical reasons.⁶ Instead, we implement recursion on top of the flat structure: we allow fillers to refer to other slots or their sequence. This effectively creates non-terminal fillers (and potentially slots) within the structure and allows us to build a kind of tree structure. In this way, we can define components grouping alternative multi-token variants.

Since both slots and fillers can also be typed, we can easily differentiate terminal and non-terminal slots (and fillers) of different types. This allows us to define additional virtual structural relations among the terminal tokens such as constituency structures for potential syntactic analysis. A side effect of this “broken” virtual recursion is thus the possibility to define multiple alternative (full or partial) tree structures of the core terminal slots, with all the obvious advantages and disadvantages.⁷

More complex dependencies have also been already registered, for example, several optional components which may either occur exclusively, but not all at the same time, or which must actually either appear all together, or not at all. Another type is represented by example (10) (cf. §3.2.1), showing a variable component used repeatedly. Some of these could be (in theory) easily marked at the syntactic level, but as explained above, we can currently only rely on the surface form (with morphological analysis at its best) and therefore we need more primitive methods to group, relate and classify some slots using additional dedicated supporting features to give proper hints to the parser.

As mentioned in §3.2.6, the creativity (or lack of knowledge) of language users may eventually go far beyond the bounds of any common variability and the MWE may be modified or reduced up to the point where it is just barely recognizable as the original MWE, so that we call it a *fragment*. For this purpose, we add another special feature for each more complex entry: the minimal list(s) of the necessary components which must necessarily occur in the text in order to make an association with the original MWE possible at all.

⁶Indexing, querying and processing recursive data structures is still a demanding task, not very well and efficiently supported by the current database and search engines.

⁷Among the advantages: multi-purpose or multi-theory use and multi-dimensionality of the core database; disadvantages: additional complex requirements on consistency and validity management, need for interpretation and filtering of the basic data on higher application levels. Querying the structure of the MWEs would also be rather difficult to implement, but this functionality is currently not needed.

4.3 Implementation

The database has been implemented as a part of a more generic database of corpus annotation units, sharing a common infrastructure and principles. Elastic-Search⁸ is used as back-end engine for searching and storing the entries in the form of JSON documents. A data model written in Python is used as an intermediate abstraction, providing a generic API.

The latest front-end user interface is designed using ReactJS.⁹ It uses the API and metadata about all defined object types (a kind of *configuration* also managed by the API) to create a customized and highly configurable user interface on the fly.

4.3.1 Populating the database with entries

To populate the LEMUR database with entries, we use an automatic conversion from the FRANTALEX lexicon, which contains lemmas and tags describing the syntactic type of the lemmas. Lemmas in FRANTALEX are divided into individual variants, for example, *jít přes čáru* ‘to cross the line’ and *být přes čáru* ‘to be beyond the line’, whereas in LEMUR we group these variants under one lemma. Also, tags for syntactic types are converted into lemma descriptions. Syntactic structures are generated using a dependency parser. After a manual check they are automatically converted to constituent tree structures. The rest of the information about each lemma has to be added manually.

The FRANTALEX lexicon consists of about 49,000 lemmas, of which about half have been transferred. LEMUR contains about 16,000 lemmas, but these include grouped lemmas from the original lexicon. A test corpus, which corresponds to the SYN2020 corpus (Jelínek et al. 2021), is annotated with more than 1.3 million collocations from the FRANTALEX lexicon and more than 722,000 collocations from the LEMUR database.

4.3.2 User interface

The user interface shows all important information about a lexical entry (its components and features) in a form suitable for human readers. Figure 2 shows the lemma *mazat někomu med kolem huby* (lit. ‘spread honey around someone’s gob’) ‘butter somebody up’. Individual words, fillers, fill the numbered slots, where some positions can be occupied by several synonyms, variant fillers. For instance,

⁸<https://www.elastic.co>

⁹<https://reactjs.org>

slot [1] is filled with the variant verb fillers *mazat* and *namazat* ‘smear’, slot [5] is filled with the variant noun fillers *huby* ‘gob’, *pusy* ‘mouth’ and *úst* ‘mouth’. Below the lemma, definitions and examples from the corpus are given, as shown in Figure 3. This is followed by an option to search for examples in the corpus.

Figure 2: MWE lemma in user interface

Definition and examples

definition lichotit někomu, aby vyhověl našemu přání, přistoupil na naše požadavky

SČFI definition mazat někomu med kolem huby/úst (Čl. vůči zvl. vlivnému druhému ve snaze si ho získat, nepohněvat ap. při líčení jeho role, zásluhy ap.) zveličovat (uváděním samých kladů) pozitivní úlohu někoho a neupřímně mu tak lichotit.

examples (1) ... a pak se předháněli, kdo mi namaže víc medu kolem huby.

Corpus search

Corpus query in corpus SYN2020lemur ▾

Figure 3: Definitions, examples and search

The interface also dynamically generates charts representing syntactic structures of the MWE from its flat list of slots and their fillers and links (relations) between them. In Figures 4 and 5 we show the dependency and constituent structures of the phrase *bojovat pro čest a slávu* ‘fight for honour and glory’ with all the lexical variants in place of the verb as well as the preposition.

In these charts, blue nodes represent terminal slots of the MWE indicating also their actual possible fillers (for example the variable choice of prepositions *pro*, *za* and *o* in the slot [2]). The dark yellow slots represent the (non-terminal) phrase nodes in the constituent structure. The light yellow non-terminal slot [V] represents the verb, which may be realized by three different types of verbs: (1) simple non-reflexive verbs (*bojovat* ‘fight’, *zápolit* ‘compete, wrestle’, etc.), (2) reflexive verbs using the accusative reflexive pronoun *se* (*bít se* ‘struggle, wrestle’, *rvát se* ‘brawl’) and (3) a reflexive verb using the dative reflexive pronoun

si (*zahrát si* ‘play, act the part of’). Since the latter two types consist of two tokens, a simple list of fillers within a singular terminal slot would not be sufficient. Therefore, the verb slot is defined as a non-terminal *variant-slot*, which branches both charts into three alternative sub-trees numbered by the respective fillers 1, 2 and 3 (shown as small elliptical yellow nodes).¹⁰

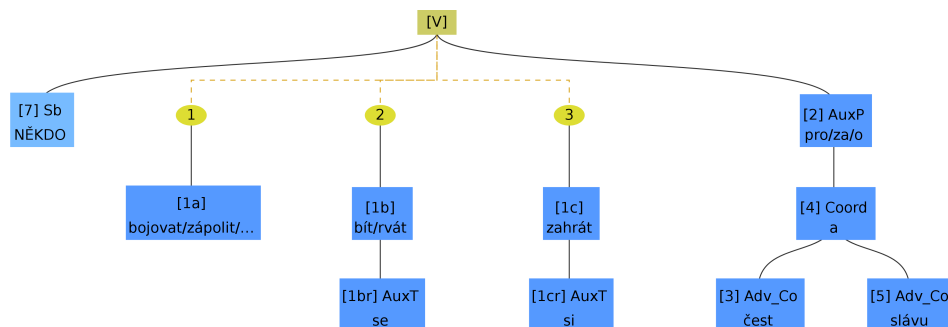


Figure 4: Dependency structure with multiple choices

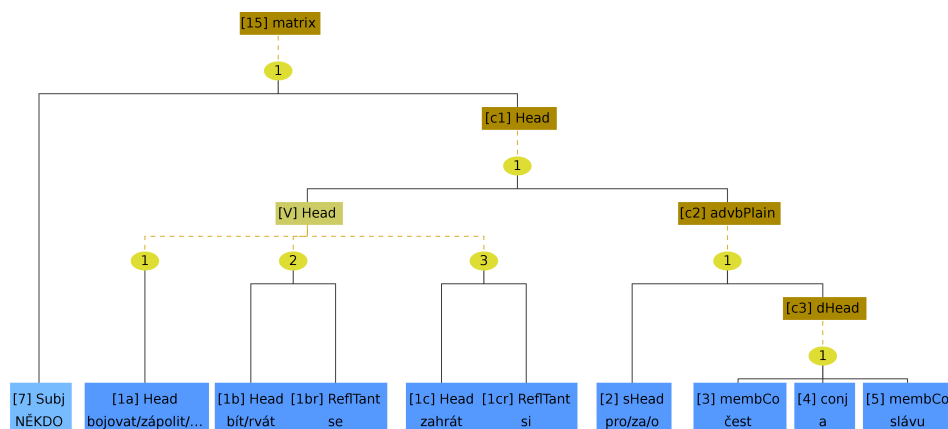


Figure 5: Constituent structure with multiple choices

In addition to the browsing mode, the database also allows editing of individual entries. In the editing mode, there is more information available that is not

¹⁰In Figure 5 (constituent structure), the fillers of all the non-terminal slots are shown as elliptical yellow nodes purely for consistency reasons, despite the fact that there is otherwise always just one filler in each of the slots and therefore no other case of branching. Terminal (single token) slots do not have their fillers branched out externally in order to keep the tree as compact as possible. For other caveats concerning the visualizations see Vondříčka (2019).

normally displayed, but can be queried when searching the lexicon. For example, various grammatical constraints, such as the occurrence of the verbal MWE only in the active voice, or only in the singular, or only in the 3rd person, etc., are expressed by constraints on the morphological tag or on the “verbtage”, a positional attribute which describes the properties of the entire verb form (whether simple or compound) such as person, voice or tense (see Jelínek et al. 2021). In Figure 6 we see the collocation *vyjít najevo* ‘come to light’, where we use a verbtage expressed by a regular expression to set the constraint that the verb can occur only in the 3rd person, in the active voice, or in the infinitive.

slot 1	slot 2
lemma: vyjít	lemma: najevo
tag: V	tag: D
verbtage: V.A3.. VFA---	

Figure 6: Morphological constraints on a MWE member

5 Practical use of the LEMUR lexicon

The lexicon can be used as a standard phraseological lexicon, but it can also be used in corpus annotation and it also has potential applications in NLP.

5.1 Annotation of MWEs in corpora and linking with the lexicon

Occurrences of MWEs in text corpora are identified and the corpus annotation is extended by the MWE lemma and type, assigned as new attributes of every token recognized as part of a MWE (in addition to the standard annotation of individual words in terms of POS tags and lemmas). Corpus users can then search for MWEs by their MWE lemma (if they know it) or they can combine various types of linguistic annotation in one search, such as a verb in imperative which is a part of a syntactically idiomatic MWE or any form of the noun *holub* ‘pigeon’ being part of a MWE. Using the Corpus Query Language in the KonText search environment (Machálek 2020) of the Czech National Corpus, the latter query would be specified as `[lemma="holub" & mwe_lemma!=""]` (`mwe_lemma` is not empty, i.e. the token with the lemma *holub* is part of an identified MWE). The user would thus find several MWEs in their context such as *pečení holubi lítají do huby* (lit. ‘roasted pigeons fly into the mouth’) ‘expectation of profit without effort’ or *točit se jako holub na báni* (lit. ‘to turn around as a pigeon on a temple dome’) ‘to turn around constantly’.

Each MWE occurrence in the corpus is linked to the corresponding lexical entry in the lexicon, so that the corpus user can consult the lexicon directly, see Figure 7 (mwe_lemmas are shown in bold characters after slashes). In the opposite direction, it is possible to view occurrences of a given MWE in a corpus when browsing the MWE lexicon, as shown above in Figure 3.

The screenshot shows a concordance search interface. The main text contains several instances of the MWE 'holub' and its variants, which are highlighted in blue. A pop-up window titled '"holub"' is open over one of the instances. The pop-up window contains the following information:

- A link to the LEMUR database: [točit se jako holub na báni](#)
- A search bar with the text 'Přaha 1' and a search icon.
- A list of search results, including 'Přaha 1', 'Předu se dostane', 'na břeh a v doprovodu po', 'on se vždycky opravdu', 'malou flotilu člunů, která', 'ušku, až budete projíždět', 'oruju', 'poslaly spojenecké armá', 's Diamond – je', 'vu neprosívá nadměrné š', 'h společenských akcí: je'.

Figure 7: Corpus concordance linked to MWE lexicon

5.2 Use of the lexicon in POS tagging and parsing

A morphological tagger may use a module that identifies some frequent MWEs in order to decide about the most likely tag using the knowledge of such MWEs rather than general linguistic rules or a stochastic model unaware of these phenomena (Hnátková & Petkevič 2017). Since MWEs are sometimes morphologically or syntactically irregular as in (36), their identification, including tagging with a special module, helps to increase the tagging accuracy of the whole corpus. For instance, in (43) the general morphosyntactic rules of Czech cannot fully disambiguate case and number of the noun *bratrství* ‘brotherhood’ (following the preposition *na* ‘on’, requiring accusative or locative, the noun *bratrství* can be interpreted as ACC.SG, LOC.SG or ACC.PL), whereas the morphologically fully disambiguated entry *připít na bratrství* helps to disambiguate this MWE within a sentence as indicated:

- (43) připít na bratrství
 drink on.ACC-VAL brotherhood.SG.ACC
 ‘raise one’s glass to brotherhood’

The proverb in (44) includes *štěstí* ‘good luck’, a highly syncretic noun form, whose interpretations are difficult to disambiguate without a MWE lexicon listing the disambiguated morphological categories. Even in this context, but without the knowledge of the proverb, the ambiguous form *štěstí* ‘good luck’ can mistakenly be parsed as dative, modified by *odvážnému* ‘to the brave’.

- (44) Odvážnému štěstí přeje.
 brave.M.SG.DAT good luck.NOM favour.3SG.PRS
 ‘Fortune favours the brave.’

5.3 Use of the lexicon in parsing

The lexicon of MWEs contains information about the syntactic structure of each MWE. In parsing, this information can be used to automatically correct the syntactic annotation of MWEs by comparing the annotation made by the parser with the annotation specified in the lexicon for each identified MWE. If they differ, the automatic annotation can be replaced with the annotation from the lexicon, unless this would result in an overall incorrect structure, such as a looped tree, in which case the correction is not performed.

As an example, consider a simple MWE type: a noun followed by an adjective. This is a typical structure of Czech terms, for example *anděl strážný* (lit. ‘angel guardian’) ‘guardian angel’, *kudlanka nábožná* (lit. ‘mantis devout’) ‘praying mantis’, *kyselina sírová* (lit. ‘acid sulphuric’) ‘sulphuric acid’, etc. However, apart from terms, the typical word order in Czech is adjective–noun. The parser cannot acquire sufficient “knowledge” of Czech terms from the limited training data, and even the use of methods based on word embeddings (creating a mathematical representation of words using extensive “raw” language data, see Mikolov et al. 2013) does not completely remove this handicap. When a term of the noun-adjective type is followed by another noun, or by an adjective and a noun, the parser decides in 58% of cases that the adjective pre-modifies the noun to its right, sometimes even when the adjective cannot agree with the following noun in number, gender or case, as in *představený kláštera Matky Boží řádu trapistů* ‘the abbot of the monastery of the Mother of God of the Trappists’, including the term *Matka Boží* ‘Mother of God’, where the parser identified the adjective *Boží* ‘of God’, ‘divine’ as a modifier of the following noun *řádu* ‘order’, instead of the

preceding noun *Matky* ‘mother’. By providing the correct syntactic structure for *Matka Boží*, the lexicon could be used to rectify this error.

The parser used for syntactic annotation is based on neural networks (Ma et al. 2018) and trained on the data of the Prague Dependency Treebank (Hajič et al. 2018). Its overall results reach the state of the art but it struggles to correctly parse MWEs with unusual syntactic structures.

Experiments in correcting syntactic annotation were performed in the past (Jelínek 2019): syntactic structures of MWEs identified in corpora were checked against the syntactic structures exported from the MWE lexicon. When they differed, the (supposedly erroneous) structures were replaced by the structures from the MWE lexicon. The whole sentence was then checked to make sure an incorrect sentence structure did not result from this intervention. Manual analysis of the results showed that using the information from the lexicon, syntactic annotation was corrected in 88% of the identified syntactic annotation errors for MWEs, while in only 2% of the cases was an error introduced into the annotation by the intervention. Overall, however, the number of interventions was small (partly due to the relatively low number of MWEs in the lexicon at the time of the experiment) and the overall success rate of the syntactic annotation was almost unaffected by the experiment (less than 1 word per 100,000 was corrected in this way). However, there are now significantly more MWEs in the lexicon, so we consider applying the module for the automatic correction of MWE parses in the next syntactically annotated corpus due to be released in 2025. This will still mean a relatively small improvement in the overall success rate, but we expect that the syntactic annotation of MWEs will improve noticeably, especially since some structures in MWEs are really unusual and thus unmanageable for the parser. This has not been tested yet, however.

6 Conclusion

To answer the need for a lexicon of Czech MWEs, we designed and implemented a lexical database, coping with the variability and structure of multiword lexemes. To achieve that, lexical entries support descriptions from a number of angles. Thus, each entry specifies aspects such as the MWE’s lemma, definition, examples, style, syntactic structure, idiomaticity and variability.

Following the taxonomy proposed by Baldwin & Kim (2010) and used in the PARSEME project, we use multiple types of both idiomaticity and variability, i.a. lexical, morphological or syntactic. While the types of idiomaticity describe the MWE’s inherent properties, lexical specifications of variability describe the

MWE's behaviour in language use. This concerns the cross-linguistically common phenomena of internal modification (insertion) and the use of MWE fragments (omission).

In addition to its use in a standard way for lexical lookup, the lexicon can also be used as a resource for various NLP tools, such as taggers or parsers. Moreover, lexical entries can be linked with occurrences of the multiword lexemes in a corpus, supporting both lexical lookup and corpus search directly from the corpus or the lexicon, respectively. There are also plans for LEMUR to be linked with an emerging standard reference lexicon of Czech: the *Academic dictionary of Contemporary Czech* (Kochová & Opavská 2016a,b).¹¹

Last but not least, the lexicon is being extended by adding new entries or by specifying additional features within existing entries. This (to a large extent manual) effort gradually alleviates the problem of insufficient coverage: the current number of tokens at the time of writing approaches 16,000, while the number of MWE occurrences identified and annotated using the FRANTALEX lexicon in the SYN corpus release 11 is about 49,000. However, we need a strategy for further expanding the lexicon. The lacunae that come up most often in real texts deserve to be filled first, thus helping to reach a better coverage with least efforts.

In the near future, we will add all FRANTALEX lemmas to LEMUR and use the LEMUR database to annotate a new experimental version of SYN corpus (Hnátková et al. 2014). This corpus will be accessible to interested lexicographers and linguists. The feedback they provide will be valuable for the further development of the lexicon.

The still insufficient coverage aside, we believe that LEMUR is built on solid foundations and hope that it turns out to be a useful resource for many purposes. Eventually, its design and structure may serve also other languages than Czech.

Abbreviations

ACC-VAL	valency (required) accusative	MWE	multiword expression
AdjP	adjective phrase	NP	noun phrase
AdvP	adverbial phrase	NLP	natural language processing
GER	German	PASSP	passive participle
ENG	English	PP	prepositional phrase
LAT	Latin	VP	verb phrase

¹¹<https://slovníkcestiny.cz/>

Acknowledgements

Thanks to the reviewers as well as the editors for their valuable comments that helped to improve and enrich this paper.

The work on this paper was supported by the grant *Czech National Corpus* LM2023044 under Large Research, Development and Innovation Infrastructures program, and by the grant *Multiword Units for Digital Learning* TQ01000177 under TA ČR VS SIGMA - DC3 program.

References

- Aijmer, Karin. 1996. *Conversational routines in English: Convention and creativity*. London: Routledge.
- Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Fred J. Damerau & Nitin Indurkha (eds.), *Handbook of natural language processing*, 2nd edn., 267–292. Boca Raton, FL: CRC Press, Taylor & Francis Group.
- Bozděchová, Ivana. 2007. Teorie terminologie v historických a obsahových proměnách. In *Sborník příspěvků věnovaných profesorce PhDr. Marii Čechové, DrSc.* 65–74. Univerzita J. E. Purkyně, Ústí nad Labem.
- Burger, Harald, Dmitri Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.). 2007. *Phraseology: An international handbook of contemporary research*. Berlin: Walter de Gruyter.
- Čechová, Marie. 2011. *Čeština: Řeč a jazyk*. Praha: SPN.
- Čermák, František. 2007. *Czech and general phraseology*. Prague: Karolinum.
- Čermák, František. 2016. Frazologie a idiomatika. In Petr Karlík, Marek Nekula & Jana Pleskalová (eds.), *Nový encyklopedický slovník češtiny*, 1st edn., 530–532. Praha: Nakladatelství Lidové noviny. <https://www.czechency.org/slovník/FRAZEOLOGIE%20A%20IDIOMATIKA>.
- Čermák, František et al. 1983–2009. *Slovník české frazeologie a idiomatiky (SČFI)*, vol. 1–4. Praha: Academia/Leda.
- Colson, Jean-Pierre. 2017. The IdiomSearch experiment: Extracting phraseology from a probabilistic network of constructions. In Ruslan Mitkov (ed.), *Computational and corpus-based phraseology*, 16–28. Cham: Springer.
- Coulmas, Florian (ed.). 1981. *Conversational routine: Explorations in standardized communication situations and prepatterned speech*. The Hague: Mouton.
- Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína

- Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová & Zdeněk Žabokrtský. 2018. *Prague Dependency Treebank 3.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-2621>.
- Hnátková, Milena. 2002. Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a Slovesnost* 63(2). 117–126. <http://sas.ujc.cas.cz/archiv.php?art=4064>.
- Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová & Pavel Vondříčka. 2017. Eye of a needle in a haystack. In Ruslan Mitkov (ed.), *Computational and corpus-based phraseology*, 160–175. Cham: Springer. DOI: 10.1007/978-3-319-69805-2_12.
- Hnátková, Milena, Michal Křen, Pavel Procházka & Hana Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 160–164. Reykjavík: ELRA. <https://aclanthology.org/L14-1267/>.
- Hnátková, Milena & Vladimír Petkevič. 2017. Morphological disambiguation of multiword expressions and its impact on the disambiguation of their environment in a sentence. *Jazykovedný Časopis* 68(2). 145–155. DOI: 10.1515/jazcas-2017-0025.
- Jelínek, Tomáš. 2019. Using a database of multiword expressions in dependency parsing. In Kamil Ekštejn (ed.), *Text, speech, and dialogue: 22nd international conference*, 19–31. Cham: Springer. DOI: 10.1007/978-3-030-27947-9_2.
- Jelínek, Tomáš. 2020. Multi-word lexical units with repetition of lexemes in Czech and identification of their variants. In Joanna Szerszunowicz & Martyna Awier (eds.), *Reproducible multiword expressions from a theoretical and empirical perspective*, 141–153. Białystok: University of Białystok. <http://hdl.handle.net/11320/11351>.
- Jelínek, Tomáš, Marie Kopřivová, Vladimír Petkevič & Hana Skoumalová. 2018. Variabilita českých frazémů v úzu. *Časopis pro moderní filologii* 100(2). 151–175. <https://casopispromodernifilologii.ff.cuni.cz/magazin/2018-100-2-2/>.
- Jelínek, Tomáš, Jan Křivan, Vladimír Petkevič, Hana Skoumalová & Jana Šindlerová. 2021. Syn2020: A new corpus of Czech with an innovated annotation. In Kamil Ekštejn, František Pártl & Miloslav Konopík (eds.), *Text, speech, and dialogue*, 48–59. Cham: Springer. DOI: 10.1007/978-3-030-83527-9_4.
- Klégr, Aleš. 2016. Lexikální kolokace: Základní přehled o vývoji pojetí. *Časopis pro moderní filologii* 98(1). 95–103. <http://hdl.handle.net/20.500.11956/96860>.

- Kochová, Pavla & Zdeňka Opavská. 2016a. Akademický slovník současné češtiny: Z přípravy Akademického slovníku současné češtiny. *Naše řeč* 99(2). 57–83. <https://ujc.avcr.cz/miranda2/export/sitesavcr/ujc/zakladni-informace/pracovnici/files/KochovaOpavskaASSC.pdf>.
- Kochová, Pavla & Zdeňka Opavská (eds.). 2016b. *Kapitoly z koncepce: Akademického slovníku současné češtiny*. Praha: Ústav pro jazyk český AV ČR.
- Kopřivová, Marie & Milena Hnátková. 2014. From dictionary to corpus. In Vida Jesenšek & Peter Grzybek (eds.), *Phraseology in dictionaries and corpora*, 155–168. Maribor: Filozofska fakulteta Maribor.
- Kovářiková, Dominika. 2017. *Kvantitativní charakteristiky termínů*. Praha: Nakladatelství Lidové noviny – Český národní korpus.
- Kovářiková, Dominika & Oleg Kovářik. 2019. Automatic identification of academic phrases for Czech. In Gloria Corpas Pastor & Ruslan Mitkov (eds.), *Computational and corpus-based phraseology (EUROPHRAS 2019)* (Lecture Notes in Computer Science 11755). Cham: Springer. DOI: 10.1007/978-3-030-30135-4_17.
- Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.
- Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.
- Lopatková, Markéta, Václava Kettnerová, Eduard Bejček, Karolína Skwarska & Zdeněk Žabokrtský. 2014. *VALLEX 2.6.3: Valency lexicon of Czech verbs*. Prague: Karolinum Press.
- Ma, Xuezhe, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig & Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics*, 1403–1414. Melbourne: Association for Computational Linguistics. DOI: 10.18653/v1/P18-1130.
- Machálek, Tomáš. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the twelfth Language Resources and Evaluation conference*, 7003–7008. Marseille: ELRA. <https://aclanthology.org/2020.lrec-1.865>.

- Martins, André, Miguel Almeida & Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, 617–622. Sophia: Association for Computational Linguistics. <https://aclanthology.org/P13-2109/>.
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*. DOI: 10.48550/arxiv.1301.3781.
- Moon, Rosamund. 2007. Corpus linguistic approaches with English corpora. In Harald Burger, Dmitri Dobrovolskij, Peter Kühn & Neal R. Norrick (eds.), *Phraseology: An international handbook of contemporary research*, 1045–1059. Berlin: Walter de Gruyter.
- Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.
- Pasquer, Caroline, Agata Savary, Jean-Yves Antoine & Carlos Ramisch. 2018. Towards a variability measure for multiword expressions. In Marilyn Walker, Heng Ji & Amanda Stent (eds.), *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, Volume 2 (Short Papers)*, 426–432. New Orleans, LA: Association for Computational Linguistics. DOI: 10.18653/v1/N18-2068.
- Popovičová, Snežana. 2020. *Česká a srbská frazeologie: Na cestě ke dvojjazyčnému frazeologickému slovníku*. Praha: Karolinum.
- Przepiórkowski, Adam, Jan Hajič, Elżbieta Hajnicz & Zdeňka Urešová. 2017. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography* 30(1). 1–38. DOI: 10.1093/ijl/ecv048.
- Rosen, Alexandr & Hana Skoumalová. 2018. No way to have your say out of the frame: Specifying valency of multi-word expressions. *Prace Filologiczne* 2018(72). 301–320. <https://www.ceeol.com/search/article-detail?id=732446>.
- Schulte im Walde, Sabine. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 269–308. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998645.
- Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 35–68. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579035.

- Temmerman, Rita. 2000. *Toward new ways of terminology description: The sociocognitive approach*. Amsterdam, Philadelphia: John Benjamins.
- Urešová, Zdeňka. 2009. Building the PDT-VALLEX valency lexicon. In *On-line proceedings of the fifth corpus linguistics conference*. University of Liverpool. <https://ufal.mff.cuni.cz/~uresova/web.pdf/2012-CLC-Building%20the%20PDT-VALLEX.pdf>.
- Vondříčka, Pavel. 2014. *Formalized contrastive lexical description: A framework for bilingual dictionaries*. München: LINCOM.
- Vondříčka, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112. 83–101. <https://ufal.mff.cuni.cz/pbml/112/art-vondricka.pdf>.

Chapter 2

Description of Pomak within IDION: Challenges in the representation of verb multiword expressions

👤 Stella Markantonatou^a, 👤 Nikolaos T. Kokkas^b,
👤 Panagiotis G. Krimpas^b, 👤 Ana O. Chiril^a, Dimitrios
Karamatskos^a, Nicolaos Valeontis^a & 👤 George Pavlidis^a

^aInstitute for Language and Speech Processing, ATHENA Research Center, Greece ^bDemocritus University of Thrace, Greece

Pomak is a non-standardised, endangered language variety of the East South Slavic dialect continuum. This article presents an online resource of 165 Pomak verbal multiword expressions collected via fieldwork. The resource has been developed with IDION, which is a web-based environment for the documentation of a wide range of multiword properties. The following information is encoded in this resource: lemma form of the expression, variants (if attested), definition in Pomak and translation in other languages, gloss, usage examples for 60 verb multiword expressions, morphosyntactic analysis in the Universal Dependencies framework as well as certain lexical relations among multiword expressions and verb alternations (if attested). Observations on the collected material that are not encoded in the Pomak edition of IDION but are presented in the article concern the types of verbal multiword expressions found in the data (light verb constructions, idioms) and the occurrence of very similar expressions in Modern Greek. The contents of Pomak-IDION are openly available; they belong to a set of resources of Pomak (corpus, morphological and syntactic models, embeddings, lexica) that have been developed as a case study of the Philotis project, which provides technological support for the documentation of living languages.



1 Introduction

This article presents a freely available online resource¹ that documents aspects of Pomak verbal multiword expressions (henceforth VMWEs). The resource, which will be referred to as Pomak-IDION, is intended to be useful to human users and Natural Language Processing (henceforth NLP) practitioners.

Pomak-IDION is a rare resource in a world where VMWE databases of endangered languages are sparse. Piirainen (2005), who has offered a very precise picture of idiom research in Europe, noted that idiom data existed for some standard European languages. She reported that minor languages and dialects were completely ignored (with a couple of exceptions). Of course, progress has been made over the years; however, databases with detailed information on idiom data for (not only European) endangered languages are still really few. Even the term “less resourced” language does not really describe the situation of endangered languages such as Pomak. An example is the report of Ní Loingsigh & Ó Raghallaigh (2016) on the development of an idiom database for Irish, which is a less-resourced European language in this respect. However, even this database draws on republished substantial work. On the contrary, there is no republished work on Pomak idioms.

Pomak is a living, endangered and non-standardised East South Slavic language variety with few written resources in various scripts. Pomak-IDION belongs to a set of state-of-the-art resources for this language (lexica, corpus, treebank, morphological and syntactic models, embeddings) that have been developed in the framework of Philotis. The Philotis project² has developed an infrastructure to facilitate the development of state-of-the-art NLP resources of living languages. The Pomak treebank has been annotated according to the Universal Dependencies formalism (henceforth UD) (Markantonatou et al. 2023).³

165 Pomak VMWEs have been collected via fieldwork. Both idioms and light verb constructions (henceforth LVCs) have been identified in this material. Several Pomak VMWEs have literal equivalents in Modern Greek. This fact suggests the presence of contact phenomena since trilingualism (Pomak, Greek, Turkish) is widespread in the Pomak community; native speakers of Greek, on the other hand, rarely speak Pomak.

To the best of our knowledge, this is the first systematic encoding of Pomak VMWEs that can be useful both to the human user and to NLP practitioners. Considerable field work was required for this task since, although usages of the

¹<https://pomak.idion.athenarc.gr/admin>

²<https://philotis.athenarc.gr/>

³<https://universaldependencies.org/>

VMWEs abound in everyday speech, they are extremely rare in the little available Pomak textual legacy. Lexical semantic relations among VMWEs, such as synonymy, are even more difficult to identify in the available corpora. We were fortunate enough to enjoy the cooperation of the Pomak community who embraced this effort and offered oral evidence.

We begin this discussion by introducing the Pomak language variety. In §2 we provide information about Pomak and the existing resources: §2.1 describes the corpus of Pomak and §2.2 the script and the orthography used in the corpus. The same orthography and script have been used in Pomak-IDION. Basic information about Pomak morphology and syntax is presented in §2.3 and §2.4 respectively and the UD Pomak treebank is introduced. §3 describes the collection of linguistic material about Pomak VMWEs.

Next, we present some observations on the collected material. §4 summarises the syntactic patterns observed in the collected material as sequences of Part of Speech (henceforth PoS) UD tags. In the collected material, both LVCs (§4.1) and idioms (§4.2) were identified; possible manifestations of contact phenomena between Pomak and Modern Greek are also addressed in these sections. More material about Pomak LVCs and idioms is provided in Appendix A and Appendix B respectively.

The Pomak VMWEs were encoded with the web-based IDION database which we introduce in §5. In this section, we briefly discuss issues to which state-of-the-art databases of VMWEs have to provide a response. Then, in light of this discussion, we explain our choices regarding the basic design principles of IDION.

There are two interfaces to the database: one available to users who access the database only to look up a VMWE (henceforth external users) and one used by registered linguists who document the VMWEs (henceforth encoders). Here, we present the interface available to external users. §6 is divided into subsections each describing the searches that are available with Pomak data. Searches supported by fuzzy matching retrieve VMWEs in lemma form (§6.1). Once the desired VMWE has been identified, other searches are available: variants (if any has been attested), gloss, translations (§5.1), usage examples and their translations (§6.2), morphosyntactic analysis of the variants and the usage examples in the UD framework (§6.3), and lexical relations among VMWEs (some of them of semantic nature), if attested (§6.4).

2 About Pomak

Pomak (endonym: Pomácky, Pomácko, Pomácku or other dialectal variants) is a non-standardised East South Slavic language variety. Apart from Greece, it is spoken in parts of Bulgaria and East Thrace (Türkiye) and in the places of Pomak diaspora (Constantinides 2007: 35). In Greece, it is spoken by about 35,000 people inhabiting the Rhodope Mountain area mainly (Adamou & Fanciullo 2018). The Pomak dialect continuum has been influenced by Greek and Turkish due to extensive bilingualism or trilingualism.

Pomak scores low on all six factors of language vitality and endangerment proposed by UNESCO (Brenzinger et al. 2003): there is little written legacy with only symbolic significance for the speakers of Pomak, the language is not taught at school, it is used mainly in family settings, and the dominant languages, namely Greek and Turkish, begin to penetrate family settings.

2.1 Textual resources of Pomak

Sporadic transcriptions and recordings of Pomak folk songs and tales have been published over the last 80 years (Theocharides 1996a,b) as well as a very few modern texts; these mostly include journalistic texts, translations from Greek and English into Pomak (Karahóga 2017), and material for teaching Pomak to Greeks as a second language (Kokkas 2004). In addition, descriptive works on Pomak morphology and grammar have been published (Papadimitriou 2013, 2008, Sandry 2013). Selected parts of this textual material have been included in a corpus of about 140,000 words, which will be made available for research purposes by Philotis. The corpus is presented here because it is the largest searchable collection of texts in Pomak and has been used as a source of VMWE instances while developing Pomak-IDION.

Table 1 shows the text genres included in the corpus and the size of the respective texts in words. Where possible, the geographical origin of the texts is also given as a hint to the Pomak variant appearing in the text.

The morphological and syntactic analysis adopted in this work draws on the approach to Pomak language that was developed in the Philotis project and is outlined in Karahóga et al. (2022) and Markantonatou et al. (2023). A Pomak treebank has been made available on the UD treebank repository along with the relevant detailed documentation.⁴

⁴<https://universaldependencies.org/qpm/index.html>

Table 1: Pomak corpus: Type, size and geographical origins of texts.

Text types	Words	Geographical origins
Folk tales	43,817	Aimonio, Glafki, Dimario, Echinós Myki, Pachni, Oreó
Language description	19,524	mixed
Journalism	25,236	Myki
Translations into Pomak	24,208	Myki, Pachni
Folk songs	18,434	mixed
Proverbs	550	mixed
Other	5,325	Myki
Total	137,094	

2.2 Pomak script and orthography

A variety of scripts and orthographies have been used so far in the Pomak textual legacy, ranging from Bulgarian-based Cyrillic to Modern Greek to an English-based Latin alphabet. Homogenisation of these texts in order to form a processable corpus required the adoption of a common script and a common orthography. To this end, the Latin-based alphabet devised and proposed by Ritvan Karahóga and Panagiotis G. Krimpas (henceforth K&K alphabet), which has a language resource-oriented accented version and a non-accented all-purpose version, has been used to transliterate the corpus semi-automatically and for the documentation of Pomak VMWEs in IDION.

The K&K alphabet has been developed to satisfy the following requirements (Karahóga et al. 2022): use of Unicode to ensure portability of the alphabet, phonetic transparency, easily learned representations of sounds (ensured by the use of similar diacritics for the same articulation sounds and the absence of digraphs) and, finally, consistent spelling not affected by predictable allophony. It should be noted that the K&K alphabet is based on the Pomak variety spoken in the area of Myki but can also partially serve as an all-variety script by allowing various predictable pronunciations of the same graph depending on the variety.

The orthographic tradition of other Slavic language varieties was taken into consideration if it did not contradict distributional and phonological evidence. For instance, certain interrogative, indefinite and negative pronouns, conjunctions and adverbs are spelled as a single word in most Slavic languages but, in the adopted Pomak orthography, are spelled as two words, e.g., *at kak* for *atkák*

‘since’, *ní kutrí* for *níkutrí* ‘nobody’ because the first word can be independently identified as a preposition or particle, and the second as an interrogative pronoun or adverb e.g., *at* ‘from; out of’, *kak* ‘how; as; like’.

2.3 Pomak morphology at a glance

Pomak common and proper nouns, determiners, adjectives, pronouns, participles and some of the numerals are morphologically marked for gender, number, case and (in)definiteness. The opposition *Animate vs. Inanimate* is overt with the nominative case of masculine plural adjectives, participles and 3rd person plural pronouns and rarely with masculine singular nouns, where it is found as residual morphological genitive/accusative. Pomak has three genders, namely masculine, feminine and neuter, and four cases, namely nominative, dative/genitive, accusative and vocative; the morphological dative case has assumed the functions of the historical dative and genitive cases, so we speak of dative/genitive case and use a notation reminiscent of this fact in glossing the Pomak examples. With possessive determiners both the number of the possessor and of the possessed object are encoded. Like most Balkan languages, Pomak has a rich inventory of diminutive and augmentative forms of nouns, adjectives, adverbs, and certain passive participles.

Pomak is special among East South Slavic languages in that, although it uses a tripartite enclitic definite article *-s*, *-t*, *-n* (Adamou & Fanciullo 2018, Constantinides 2020, Krimpas 2020) like Macedonian, this article is of the *-s*, *-t*, *-n* rather than *-v*, *-t*, *-n* type and occurs not only with nominals, but also with deictic adverbs as a deictic and definiteness marker, denoting:

- Proximity to the speaker, e.g., *čulákos* ‘the man close to the speaker’.
- Proximity to the listener, e.g., *čulákot* ‘the man close to the listener’.
- Distance from both the speaker and the listener, e.g., *čulákon* ‘the man who is away (or out of sight) from both the speaker and the listener’.

Verbs have finite and non-finite forms. There are three types of non-finite verb forms: converbs, participles and (residual) infinitives. The residual, i.e., Proto-Slavic, infinitive forms the prohibitive imperative when following the particles *na/ne* and *namój* (sing.)/*namójte* (pl.) ‘not’, e.g., *namój barzá* ‘do not rush’. Interestingly, Pomak has another, innovative form of infinitive, which may be called *the morphologically reduplicated infinitive*. This residual infinitive of a small number of imperfective verbs is repeated to form fixed multiword expressions that

denote the continuous/monotonous/rhythmic repetition of a motion, e.g. *čúktiti čúktiti* ‘hit and hit’.

Finite verbs are always marked for mood, number and person. Verbs in the indicative mood are marked for tense, either past or present. *Som* ‘be’ is the auxiliary verb used to form perfect verb tenses and the passive voice. Future tenses are formed with the indeclinable auxiliary particle *še* ‘will’, which historically derives from the verb meaning ‘want’.

2.4 Pomak syntax at a glance

Pomak is a nominative-accusative language, where subjects are typically marked with the nominative case and objects with the accusative; in addition, some verbs select objects in the dative/genitive case. Indirect objects are marked with the dative/genitive case, which is morphologically based on the Slavic dative case. As in other Slavic languages, ethical datives abound. The strong and the weak forms of the personal pronoun may co-occur in a sentence (clitic doubling).

Markers such as *óti*, *da*, *če*, *ta* introduce subordinated clauses that function as verb dependents and markers such as *akú*, *kugá*, *pak*, *za da*, *za to*, *óti* introduce clauses that function as adverbial modifiers. There is a question particle *li*, e.g., *dojdés li* ‘do you come?’.

With respect to word order, Pomak is a primarily SVO language with rather flexible word order, given its highly inflectional nature. Adjectives typically come before the noun, although the reverse is also possible, especially for emphasis or in literary contexts. Possessives are actually datives of the unstressed (enclitic) personal pronoun. The rules governing the word order of clitics in a clause, as well as the word order within a clitic cluster, are similar to those of Bulgarian, Macedonian and Serbo-Croat: a single clitic is always the second element of its clause; multiple clitics are arranged in the following order: auxiliary > clitic-in-dative-case > clitic-in-accusative-case but, if the auxiliary is 3rd person singular, the order changes into clitic-in-dative-case > clitic-in-accusative-case > auxiliary. Pomak is a pro-drop language, which means that pronominal subjects are normally used for clarity, emphasis, or literary purposes since verb endings normally provide information about the “number” and “person” of the syntactic subject. Given that infinitives are no longer in use in Pomak (except for the residual and reduplicative infinitives mentioned above), the so-called “Balkan subjunctive” (*da* particle + finite verb in the case of Pomak) has replaced the old Slavonic infinitive (much like Modern Greek, Albanian, Romanian, Bulgarian, Macedonian and, to a lesser extent, Serbian and Bosnian).

3 Material collection

Pomak VMWEs were collected mainly through interaction with native speakers of Pomak in the framework of Philotis. The collection of VMWEs was accomplished by Nicolaos Kokkas, one of the authors, who is fluent in Pomak. The native speakers who contributed to this study represented the variants of Pomak spoken in the following villages in the region of Xanthi: Bára (Greek name Στήριγμα ‘Stírigma’), Bašájkovo (Greek name Μάνταινα ‘Mándena’, Demirǵík (Greek name Δημάριο ‘Dimáριο’), Púlevo (Greek name Προσήλιο ‘Prosilio’).

Targeted interaction with native speakers involved (recorded) interviews and collection of written material. The speakers were two men and two women of secondary and tertiary education level, whose ages ranged between 20 and 50 years. During the interviews specific VMWEs were discussed. To collect written material, during the period September–December 2022, each week the speakers received a short list of VMWEs which they discussed in their community and enriched with semantically related VMWEs, namely synonyms and antonyms⁵ (if they could identify any) and usage examples. Written material was collected in the form shown in Table 2 (<> indicates translation into the respective language, e.g. <Greek>: translation into Greek, [Pomak] any original text in Pomak and (Gloss) the gloss of the Pomak VMWE in Modern Greek or English).

Table 2: Form used to collect evidence about Pomak VMWEs.

VMWE	[Pomak]	<Greek>	<English>
Definition	[Pomak]		(Gloss)
Synonyms	[Pomak]	<Greek>	
Opposites	[Pomak]		
Examples	[Pomak]	<Greek>	<English>
ID:	Interviewed		Date:
	speaker(s):		

The forms were further filled with material from the Pomak corpus that was searched for in-context usages of the VMWEs in a variety of texts (however, little material was collected in this way). Recordings of Pomak contemporary speech obtained in 2022 provided more VMWE instances of usage. The authors of this chapter have encoded the collected material.

⁵In IDION, the term *opposites* is preferred rather than the term *antonyms* for reasons explained in §5.6

4 A closer look into Pomak VMWEs

In this section, we will take a closer look at the structure of the collected Pomak VMWEs.

In this article, we will make frequent use of the term *lexicalised components of a MWE* which was introduced in Savary et al. (2018: 94). These are the components with either fixed form or fixed lemma. Apart from the lexicalised components, VMWEs have free components that set them apart from proverbs; still, these free components are subject to strong semantic and morphosyntactic constraints. Throughout this article, the lexicalised components of the VMWEs that are used as examples are typed in bold. The slots in the VMWE that should be filled with free arguments are indicated by means of pronouns in regular script.

Based on the collected data, a set of observations have been made; these observations are not available in the existing literature on Pomak and/or on Pomak idiomaticity:

- Pomak uses LVC constructions.
- The set of light verbs identified in the Pomak data is very similar to those of other European languages (see §4.1).
- The pattern VERB+NOUN is very frequent in LVCs and in idioms (1).
- Pomak VMWEs demonstrate verb alternation phenomena (see §6.4).
- Several Pomak idioms have literal equivalents in Modern Greek. This observation could possibly contribute to a wider study of idiomaticity in the Balkan languages.

The syntactic patterns of the lexicalised components of the collected VMWEs are listed below as PoS sequences. When a literally equivalent Greek VMWE exists, this is introduced with the prefix “GE” (Greek Equivalent) next to the English translation of the Pomak VMWE. Of the syntactic patterns (1–8), (1) has been attested in both idioms and LVCs and the other patterns in idioms only. It should be noted that all these patterns are in use in non-idiomatic Pomak. Throughout this text, the infinitive is not used in the English glosses of verbs because both Pomak and Modern Greek use the verb’s 1SG.PRES.IND. form as its lemma form.

- (1) VERB + NOUN (LVCs; certain idioms)
- (2) VERB + ADJECTIVE
stánavom fukará
 become.1SG.VERB poor.ADJ.SG.NOM
 ‘I become poor’ Verb: *fukarjásavom* ‘I become poor’
- (3) VERB + ADPOSITION + NOUN
astánavom na másto
 remain.1SG.VERB at.ADP place.NOUN.SG.ACC
 ‘I die instantaneously’
- (4) VERB + ADPOSITION + ADJECTIVE + NOUN
astánavom sas atvórena ustá
 remain.1SG.VERB with.ADP open.ADJ.SG.FEM.ACC mouth.NOUN.SG.FEM.ACC
 ‘I remain speechless’ GE: *μένω με το στόμα ανοιχτό*
- (5) VERB + NOUN + ADPOSITION + NOUN
atvárem belé na glavóso
 open.1SG.VERB trouble.NOUN.ACC on.ADP head.NOUN.SG.ACC
 ‘I cause problems to myself’ GE: *βάζω μελά στο κεφάλι μου*
- (6) VERB + ADJECTIVE + ADPOSITION + NOUN
právem bannóga čórna ad sópa
 do.1SG.VERB somebody black.ADJ.ACC from.ADP beating.NOUN.SG.ACC
 ‘I beat someone hard’ GE: *κάνω μαύρο στο ξύλο κάποιον*
- (7) VERB + ADPOSITION + NOUN
klávom náeko faf óči
 put.1SG.VERB something in.ADP eye.NOUN.DEF.PL.ACC
 I crave for something
- (8) VERB + ADVERB
glódom kríve bannóga
 look.1SG.VERB away.ADV somebody
 ‘I glare at somebody’

Word order permutations can be observed in the collected material. Here one sees, e.g., that non-lexicalised variable indirect objects may come either after all the lexicalised parts of the VMWE, or immediately after the verb of the VMWE.

- (9) a. **dávom kolájene bannómu** OR **dávom bannómu kolájene**
 give.1SG eases somebody.GEN
 ‘I greet somebody’

- b. **dávom habér** bannómu OR **dávom bannómu habér**
 give.1SG news.NOUN somebody.GEN
 ‘I inform somebody’

4.1 Pomak LVCs in the collected material

LVCs were first introduced by Jespersen (1965) and since then they have attracted a lot of attention (e.g., Baldwin & Kim 2010, Laporte 2018). LVCs consist of a verb and a nominal complement, possibly introduced by a preposition. Savary et al. (2018) list a set of diagnostics for setting LVCs apart from idioms with a VERB+(PREPOSITION)+NOUN syntactic structure: the noun has one of its original senses and denotes an event or a state; the verb only contributes morphological features, such as tense, mood, person and number; the noun can head an NP containing all the syntactic arguments of the verb and denoting the same event or state as the LVC; and, the overall construction is subject to semantic and syntactic uniqueness constraints.

Here, we identify as LVCs those VERB+NOUN formations that can be replaced by (are synonymous with) verbs that are morphologically related to their noun (10); such structures seem to satisfy the LVC diagnostics listed above. Among the Pomak verbs used as light verbs are *dávom* ‘I give’, *právem* ‘I do’, *stánavom* ‘I become’, *stórevom* ‘I make’, *zímom* ‘I take’. More examples of Pomak LVCs are listed in Appendix A.

- (10) a. **dávom** **izét**
 give.1SG.VERB pain.NOUN.SG.ACC
 ‘I torture’ Verb: *izettóvom* ‘I torture’
- b. **stánavom** **fukará**
 become.1SG.VERB poor.ADJ.SG.NOM
 ‘I become poor’. Verb: *fukarjásavom* ‘I become poor’
- c. **stórevom** **izméte**
 do.1SG.VERB service.NOUN.PL.ACC
 ‘I do the housework’. Verb: *izmetóvom* ‘I serve’
- d. **zímom** **emín**
 take.1SG.VERB oath.NOUN.SG.ACC
 ‘I take an oath’. Verb *eminledísavom* ‘I take oath, I vow’

The Pomak corpus has provided some usage examples of LVCs (11):

- (11) a. Nimó ma právi rezíl.
do-not.2SG.VERB me do infamous.ADJ
'Do not humiliate me.'
- b. Čuláekon zíma karáre annók déne da
man.the take.3SG.VERB decision.NOUN.PL.ACC one day that
íde da nájde Alláha.
go.3SG.VERB that find.3SG.VERB Allah
'One day, the man makes the decision to go and find Allah.'

4.2 Idioms occurring in both Pomak and Modern Greek

In our collection of 165 Pomak VMWEs, we traced 55 VMWEs that have literal equivalents in Modern Greek. We consider two VMWEs as *literally equivalent* if they consist of translationally equivalent lexicalised parts for the same non-compositional meaning. These data may present an interesting aspect of language contact phenomena between Greek and Pomak or, perhaps, an instance of wider linguistic interactions in the Balkans or other parts of Europe (Piirainen 2005, Krimpas 2022). More VMWEs of this type are listed in Appendix B. Some pairs of equivalent Pomak and Modern Greek VMWEs are exemplified in (12).

- (12) a. i. **ablízavom si pórstovene**
lick.1SG.VERB I.PRON finger.NOUN.DEF.PL
'I find the food delicious'
- ii. GE: γλείφω τα δάχτυλά μου
glifo ta dachtila mou
lick.1SG.VERB the.ART.PL.ACC finger.NOUN.PL.ACC my
'I find the food delicious'
- b. i. **čéftom balíkoso**
chisel.1SG.VERB wound.NOUN.DEF.PL.ACC
'I open old wounds'
- ii. GE: ξύνω πληγές
ksino pliges
chisel.1SG.VERB wound.NOUN.PL.ACC
'I open old wounds'
- c. i. **klávom dvéne nógy na annó**
put.1SG.VERB two.NUM.DEF foot.NOUN.DUAL.ACC on.ADP one.NUM
amenýe
shoe.NOUN.SG.ACC
'I try to control somebody's life'

- ii. GE: βάζω τα δυο πόδια κάποιου σε ένα παπούτσι
vazo ta dio podia
 put.1SG.VERB the.ART.PL.ACC two.NUM feet.NOUN.PL.ACC
kapiou se ena papoutsi
 someone.GEN in.ADP one.NUM shoe.NOUN.SG.ACC
 ‘I try to control somebody’s life’
- d. i. **sečé mi akýlos**
 cut.3SG.VERB I.DET.GEN brain.NOUN.DEF.SG.NOM
 ‘I am intelligent’
- ii. GE: κόβει το μυαλό μου
kovi to mialo mou
 cut.3SG.VERB the.ART.SG.NOM brain.NOUN.SG.NOM my
 ‘I am intelligent’

5 Issues in VMWE documentation: The IDION approach

Modern MWE databases are expected to provide information that can be used both by people who study or use a language and in NLP (Grégoire 2010, Losnegaard et al. 2016). Gantar et al. (2018) compare seven dictionaries and NLP databases and list the MWE properties they document, namely: (i) variants (ii) definition (iii) morphology of MWE components (iv) contiguity of MWE components (v) phrase structure (vi) usage example. In what follows, we discuss these properties and how they are treated in IDION. Furthermore, we extend our discussion to additional information about VMWEs that is encoded in IDION and includes a variety of semantic properties and the full morphosyntactic description of VMWE lemmas and usage examples according to the UD framework.

IDION is a web environment for the rich documentation of MWEs. IDION allows for new editions, accessible from the same or a different site. So far, two editions have been created, one for Modern Greek VMWEs (Markantonatou et al. 2019) and one for Pomak VMWEs. The contents are available under a CC-BY-NC license.

5.1 Lemma form

In IDION, a *lemma form* of a VMWE contains:

- the components with a fixed form (fixed lexicalised components);

- the components whose lemma is fixed but whose form inflects; these are included in their lemma form or the form that best approximates the lemma convention (non-fixed lexicalised components), e.g., if the (head) verb of the VMWE appears in the second and third persons of all numbers, in the lemma form it is in the second person singular;
- the variables, such as free NPs functioning as subjects, direct/indirect objects or ethical genitives or datives of the MWE and free phrasal complements.

In addition, the lemma form takes into account the possibly fixed order of the MWE components; otherwise, it keeps the lexicalised components close to the verb. Such a typical order is the following: (lexicalised or free) subject, lexicalised verb, other lexicalised components (if any), (lexicalised or free) object. Attested usage instances of the VMWE with a different word order are separately listed in the CORPUS tab of the IDION-Pomak database (see §6.2) as manifestations of the syntactic flexibility of the VMWE.

Below, in order to better explain IDION's features we may resort to examples from Modern Greek since Pomak could provide only limited material.

5.2 Variants

Variants have to do with the lemma form of the MWEs. The lemma form is one of the two features of a MWE that have to be considered in order to create an entry in the database; meaning is the second feature. It turns out that the identity of the VMWE is established as a combination of a meaning with a non-empty set of lemma forms, the so-called variants (Vondříčka 2019). The issue of variants occurs because VMWEs are mutable entities of spoken, colloquial language. In other words, it is not the case that each VMWE lemma form corresponds to a different meaning and vice versa. For instance, all the lemma forms of the Modern Greek VMWE in (13) share the same meaning. These lemma forms are identical as regards the syntactic dependencies among the lexicalised parts that belong to content word categories, namely nouns, adjectives and verbs. In the same spirit, optional or mutually exclusive lexicalised non-content word components of the MWE may define new variants but not a new VMWE. In (13) four different lemma forms of the same VMWE result from the optionality of the article *τη* and the exclusive disjunction between *γύρω από το* and *στο*.

It should be clarified that the syntactically flexible usages of VMWEs (see §5.5) are not treated as VMWE variants in IDION.

- (13) a. GE: βάζω τη θηλειά γύρω από το λαιμό κάποιου
vazo ti thilia giro apo to lemo kapiou
 put.1SG the noose around from the neck somebody.GEN
- b. GE: βάζω θηλειά γύρω από το λαιμό κάποιου
vazo thilia giro apo to lemo kapiou
 put.1SG noose around from the neck somebody.GEN
- c. GE: βάζω τη θηλειά στο λαιμό κάποιου
vazo ti thilia sto lemo kapiou
 put.1SG the noose to.the neck somebody.GEN
- d. GE: βάζω θηλειά στο λαιμό κάποιου
vazo thilia sto lemo kapiou
 put.1SG noose to.the neck somebody.GEN
 ‘I force someone to be involved in an unpleasant situation’

A lexicalised content word component of the VMWE may appear in both the singular and the plural with no consequences for the idiomatic meaning. This situation is not exactly rare but it is unpredictable and part of the idiomatic character of a VMWE. Variation in number may induce changes to other lexicalised components of the MWE, e.g., the singular and plural lexicalised subjects in (14a) and (14b) respectively induce agreement phenomena on the (lexicalised) verbs of the respective variants of the same VMWE.

- (14) a. GE: πήρε αέρα το μυαλό κάποιου
pire aera to mialo kapiou
 take.3SG.PAST air the.SG.NOM brain.SG.NOM somebody.GEN
 ‘to get above oneself’
- b. GE: πήραν αέρα τα μυαλά κάποιου
piran aera ta miala kapiou
 take.3PL.PAST air the.PL.NOM brain.PL.NOM somebody.GEN
 ‘to get above oneself’

(15) shows the VMWE in (13) with an ethical genitive⁶ rather than a possessive one (Sailer & Markantonatou 2016). The ethical genitive alternation in (15) has to do with the morphosyntactic form of a variable and does not affect the meaning of the expression. Given this fact and the wide use of VMWEs with ethical genitives, in IDION lemma forms with an ethical genitive are listed as variants of

⁶Modern Greek: ethical genitive; Pomak: ethical dative.

the lemma forms exemplifying the other member of the alternation pair; the latter contains either an inalienable possession structure or a suitable prepositional phrase.

- (15) GE: του βάζω (τη) θηλειά [γύρω από το] / [στο] λαιμό
 tou vazo (ti) thilia [giro apo to] / [sto] lemo
 I.PRON.GEN put.ISG (the) snoose [around from the] / [to.the] neck
 ‘I force someone to be involved in an unpleasant situation’

Variants are considered a challenging feature of MWEs (Vondřička 2019, Grégoire 2010, Villavicencio et al. 2004, Skoumalová et al. 2024 [this volume]) because criteria such as the ones presented above are required to decide which forms will be listed as variants under the same MWE entry and which ones will not. No general agreement on this issue has been achieved as yet. For instance, VMWEs are often members of sets of expressions that stand in various lexical and semantic relations as discussed in §5.6. In IDION, variants are members of a set of lemma forms of one VMWE. VMWEs that differ in lexicalised content words and/or semantically define separate entries in the database. IDION allows for the encoding of sets of lemma forms because it considers these variations important for the human user and for NLP.

In developing IDION, once a first decision about a meaning and form combination is made, encoders collect as many variants and syntactically flexible usage instances as possible from corpora and/or the web. This procedure may change the original decision about the identity of the VMWE. For instance, it may arise that there are more meanings out there corresponding to the same set of forms than originally expected, or that there are forms that cannot be considered variants of the documented VMWE, for instance, because their syntactic structure cannot be reduced to the structure of the documented one. Therefore, at the heart of IDION stands the collection of usage instances of the VMWE that determines the amount and the types of information on VMWEs to be documented in IDION. It should be stressed that IDION relies on actual usage examples, preferably collected from corpora and the web. There is room for encoding the intuitions of native speakers but these examples are kept to a minimum and are marked as such. Eventually, a non-empty set of variants is collected. The longest one is chosen as the “preferred variant” and represents the VMWE.

No contracted representations are used for the lemma forms, such as representations based on regular expressions; for a comprehensive discussion on VMWE representations see Lichte et al. (2019). Since we have drawn on limited lexicographic and financial resources, we preferred to invest in the collection and study

of usage examples. Furthermore, given the current NLP technology, morphosyntactic representations of both the lemma form of the VMWE and its usage examples can be obtained, edited, and searched for structural patterns with open-source tools, thus facilitating (steps of the) encoding without any additional machinery. In addition, usage examples constitute a valuable reusable resource for model development and that was a strong motivation because IDION is designed to support NLP. Finally, human users profit from usage examples because they illustrate usage particularities that can hardly be included in the definition of the meaning of the VMWE.

5.3 Meaning, glossing, translations

In IDION, the definition of the VMWE is a short text describing the meaning of the MWE. Definitions are in the language of the VMWE and contain compositional expressions only. Because the type of arguments a VMWE supports (the variables) is an important contribution to its meaning, in the definition pronouns like ‘someone’ and ‘something’ stand for nominal complements denoting humans and non-humans, if such constraints are imposed by the VMWE.

The representative lemma form is glossed and translated into a language other than that of the VMWE. Glosses are simple with no morphological and syntactic annotation since this information is given via the UD analysis of a lemma form, which is also made available in IDION. Glosses are addressed to the human user who can complement them with the UD analysis of the lemma form. On the translation level, MWEs with an equivalent meaning are preferred when they exist in the target language.

5.4 About morphology and syntax

The morphological and syntactic analyses of MWEs are necessary both for the precise definition of the MWE form and for supporting NLP. In rule-based NLP, an important task is the development of computational lexica of MWEs enriched with the full inflectional paradigms of the entries, e.g., Savary (2009) for compounds in several languages and Al-Haj et al. (2013) for Hebrew. This requires a morphological and a syntactic description of both the language system to which the MWE belongs and the particularities of each MWE.

The morphological and syntactic representation of the MWEs must be compatible with the formal language and the framework used by the NLP tool that they will support; this raises reusability concerns. For instance, databases aimed

at supporting phrase structure-based NLP (Grégoire 2010, Vondřička 2019) employ encoding schemes that allow for non-terminal nodes. To be reused in, for instance, the UD framework, which is compatible with several popular state-of-the-art non-rule-based NLP tools and is adopted in several state-of-the-art MWE databases including IDION (Skoumalová et al. 2024 [this volume], Leseva et al. 2024 [this volume], Osenova & Simov 2024 [this volume]), these encodings have to be adapted accordingly. This is because UD uses no non-terminal nodes, has its own metalanguage for morphosyntactic annotation and the analysis is encoded in CoNLL-U.⁷ On the other hand, state-of-the-art NLP tools learn from data, so they can be possibly trained on the (adapted) inflectional paradigms of VMWEs or, alternatively, on appropriately annotated corpora of diverse and syntactically flexible usages of MWEs (Savary et al. 2019).

However, this need for many flexible usage instances proves to be hard for less-resourced languages, let alone for endangered ones. It is hard to construct corpora of spoken languages for which even a consensus on their alphabet and orthography has not yet been reached; Pomak is such an endangered language (Karahóga et al. 2022). Also, less-resourced languages with only a few corpora representing their spoken version can hardly provide syntactically flexible usages of MWEs. For instance, in the case of Modern Greek, which is a medium-resourced language according to the criteria proposed by Joshi et al. (2020), only the web (and not the published corpora) offers a reasonable amount of representative usage instances of most of the VMWEs, let alone their syntactically flexible ones.

5.5 Syntactic flexibility

The syntactic flexibility of the VMWE is documented separately with six diagnostics. Each diagnostic is exemplified with usages from the corpus. As a result, all the collected usage instances are marked for at least one syntactic phenomenon. The six diagnostics are briefly explained below:

- Subject-head verb flexibility: Can the VMWE accept different subjects? Can it appear in all persons/numbers/tenses/moods?
- Can word order variation phenomena be observed with this VMWE?
- Interpolation: Can adverbs, adjectives or even phrases occur in between the lexicalised components of the VMWE?

⁷CoNLL-U is the encoding scheme adopted by UD and the tools that process annotated corpora with the UD annotation scheme: <https://universaldependencies.org/v2/conll-u.html>

- Cliticisation of lexicalised nominal content word components.
- Passive voice: does the VMWE have both an active and a passive form?
- Ethical genitive (for Pomak, dative) alternation (see §5.2).

It has already been pointed out in §5.2 that flexible usages of the VMWE are not considered variants of the VMWE apart from the usages containing ethical genitives/datives.

5.6 Lexical (semantic) relations among VMWEs

Lexical semantic relations have found their way into state-of-the-art databases for MWEs (Leseva et al. 2024 [this volume], Giouli et al. 2024 [this volume]). In Skoumalová et al. (2024 [this volume]), the notion of “super lemma” approximates that of (several) lexical semantic relations from a different point of view. IDION documents a set of lexical (semantic) relations among VMWEs. In order for a relation to be defined, it has to be attested with a usage example. Here, we will discuss the following pairs: synonyms, opposites, Has_causative (inverse: Has_inchoative), Verb alternation that have been attested in our collection of Pomak VMWEs.

A comment is due on synonymy. Synonymy in IDION disregards stylistic differences such as +/-colloquial, +/-offensive and rather relies on a notion of *close semantic proximity* (Hüllen 2004: 39). It is well known that synonymy cannot be considered the linguistic equality relation because, in this way, synonyms would be the words or phrases capable of substituting each other in any context and such words or phrases hardly exist in any language. On the other hand, our everyday linguistic practice seems to consider synonymy a fact, e.g., when we explain the meaning of a word or a phrase using the language to which they belong (Hüllen 2004: 38).

VMWEs are also documented for opposites, that is VMWEs describing situations that cannot hold simultaneously for the same entities, e.g., one cannot at the same time be denoted by the subject of the (EN) VMWE *to kick the bucket* and the VMWE *to be alive and kicking*. Opposition in language is a multi-dimensional and much discussed phenomenon and a rich terminology has been devised for its description (Lyons 1977: 270–287). In IDION, we have chosen the term *opposites* because it seems to denote the general idea described above. We have not used the term *antonym* because it has been devised to describe a relation among gradable words (Lyons 1977).

5.7 Other relations among VMWEs

We now turn to the causative/inchoative alternation and the relation among VMWEs which in IDION is called *verb alternation relation*. Strictly speaking, the causative/inchoative and verb alternation relations are defined over verbs; VMWEs, on the other hand, are structures headed by verbs. We use these terms to describe relations among VMWEs with verb heads standing in the respective relations.

The causative/inchoative alternation has been discussed extensively in the literature. Haspelmath (1993: 90) describes the phenomenon that is defined over pairs of verbs as follows:

...it is a pair of verbs that express basically the same situations (generally a change of state, more rarely a going-on) and differ only in that the causative verb meaning includes an agent participant who causes the situation, whereas the inchoative verb meaning excludes a causing agent and presents the situation as occurring spontaneously.

He further distinguishes various morphological types of alternation, one of which is the *labile* type, where the same verb is used both in the inchoative and the causative sense.

In the literature, the term *verb alternation* has been used as a cover term for a large set of phenomena, whereby a verb supports different subcategorisation frames with relatively minor and systematic differences in meaning, such as the *spray-load* alternation and the passive voice. In IDION, a restricted use of the term *verb alternation* is made: practically, it is used for those verb alternations that have not been assigned their own label in the database, for instance, passivisation and causative/inchoative alternation have their own labels and the relevant VMWE pairs are not assigned the “verb alternation” label.

5.7.1 UD representation

In IDION, UD representations are provided for the variants and the corpus examples and offer full morphosyntactic analysis. At the moment, IDION adopts the standard UD approach according to which VMWEs are analysed in the same way as compositional structures (de Marneffe et al. 2021: 281). These UD representations are very useful to state-of-the-art NLP as training or fine-tuning material (Savary et al. 2019).

6 Pomak-IDION: The Pomak edition of IDION

In §5 we explained the main ideas regarding the documentation of VMWEs in IDION. In §1 we mentioned that IDION has two interfaces: one for encoders and one for external users. This section presents the information on Pomak VMWEs that can be retrieved from IDION.⁸ At the same time, it presents the interface for external users. A description of the interface for encoders can be found in Markantonatou et al. (2019).

The properties documented in Pomak-IDION enable the searches described in this section and are summarised in Table 3. Searching facilities were designed to conform to (i) the “what you see is what you get”, or WYSIWYG concept,⁹ and (ii) the ten heuristic criteria that describe a user-friendly interface for simplicity of use and navigation (Nielsen & Molich 1990).

Table 3: VMWE properties encoded in Pomak-IDION

1	Lemma form, definition orthographic variations	Pomak
2	Translations	English, Modern Greek
3	Codification for NLP	UD analysis (lemma form, variants)
4	Corpus	Usage examples by native speakers
5	Synonyms	Pomak VMWEs
6	Opposites	Pomak VMWEs

6.1 Fuzzy matching for VMWE retrieval

The Pomak VMWEs shown in (16) will serve as a working example.

- (16) a. *náeko mi alóknava dušo-no*
 something me.DAT unburden.3SG soul-the.ACC
 ‘something makes me feel relieved of anxiety’
- b. *alóknava mi dušá-sa*
 unburden.3SG me.DAT soul-the.ACC
 ‘I feel relieved of anxiety’

⁸It should be noted that only part of the encoding and search capabilities of IDION have been used in Pomak-IDION, since the required data, such as utterances demonstrating the syntactic flexibility of VMWEs cannot be easily obtained in the case of an under-resourced language (see discussion in §5.4).

⁹<https://www.merriam-webster.com/dictionary/WYSIWYG>

- c. **alóknava** mi **na dušó-no**
 unburden.3SG me.DAT to soul-the.ACC
 ‘I feel relieved of anxiety’

A VMWE can be retrieved with segments of its lemma form (Figure 1); this is a fuzzy matching facility that returns a, possibly empty, list of VMWEs in lemma form, each one with its definition in Pomak (Figure 2). Fuzzy matching is applied to all the variants of a VMWE; the reader may recall that the variants are listed in their lemma form and that the longest variant is used as the preferred one (see §5.1). However, few VMWEs come with variants in the Pomak edition of IDION given the way data was collected. Translations of the VMWE into other languages are accessible through the screen with the (fuzzy matching) search results.

Expression

 Insert an idiom or (consecutive) parts of it. Then select a language.
 ▼

Figure 1: Search with fuzzy matching in IDION-Pomak.

Expression	Definition
alóknava mi dušása	húbbe som
alóknava mi na dušóno	húbovo mi stánava, rahatladísavom
næko mi alóknava dušóno	næko mi stóreva hubbe

Figure 2: Searched with the string *alók* (Figure 1), IDION-Pomak returns 3 VMWEs.

When a VMWE is selected, a set of tabs pops up at the lower part of the screen. The first tab on the left provides access to the orthographic variants of the lemma form (if any exist). The second tab shows the gloss of the VMWE (Figure 3). More tabs are available and described in §6.2–§6.4.



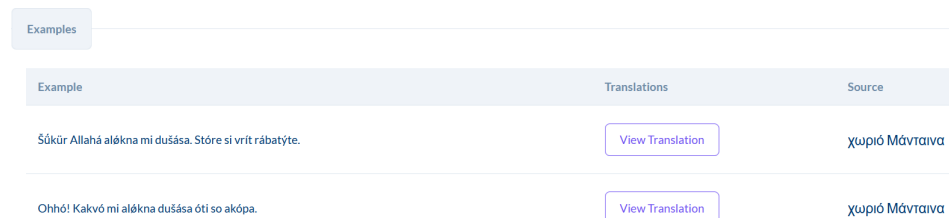
Variant: alóknava mi na dušóno

Word	Gloss
alóknava	relieves
mi	to me
na	to
dušóno	soul-the

Figure 3: Gloss of (16c)

6.2 Usage examples

The Corpus tab provides access to usage examples of the VMWE (Figure 4). For each usage example, a set of translations and the source of the example are available. The Source tab provides the name of the village of the speaker who contributed the respective usage example; for instance, in Figure 4 both usage examples have as their source the village of Mándena. Book references and URLs are normally used as sources of examples. However, the vast majority of Pomak usage examples of VMWEs were collected by means of interviews with native speakers (§3).



Example	Translations	Source
Šókkúr Allahá alóknava mi dušása. Stóre si vrit rábatýte.	View Translation	χωριό Μάντανα
Ohhó! Kakvó mi alóknava dušása óti so akópa.	View Translation	χωριό Μάντανα

Figure 4: Usage examples of (16b).

6.3 UD analysis of the lemma form

The UD-analysis tab gives the graphical format (Figure 5) and the CoNLL-U format of the analysis of the variants of the lemma form according to the UD formalism; the CoNLL-U version of the UD analysis can be viewed and downloaded through the dedicated button. The UD analysis, together with the gloss of the lemma form (Figure 3), offer detailed structural information about the VMWE. The analysis draws on the approach to Pomak morphology and syntax that has been applied on the UD Pomak treebank; this approach is outlined coarsely in §2.2, §2.3 and §2.4.

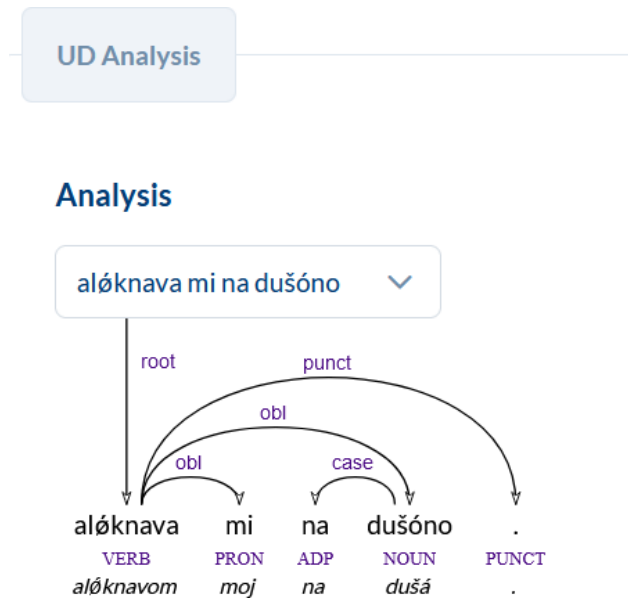


Figure 5: Graphical format of the UD analysis of (16c)

6.4 Lexical (semantic) relations: Other relations

In Figure 6 the synonyms and opposites of (16c) are given. In addition, there is a VMWE standing in the verb alternation relation with (16c).

Our data show that Pomak exemplifies the labile type of the causative/inchoative alternation (§5.6). The labels *Has_causative* and *Has_inchoative* are used to annotate pairs of VMWEs that stand in this relation. In Figure 7, the causative VMWE (16a) has the inchoative counterpart (16b). To our knowledge, this is the first time that verb alternation phenomena have been discussed for Pomak.

Open in new tab

Other Relations

Relation ↑↓	Expression
Opposite	atóžnavá mi na dušóno
Verb Alternation	násko mi alóknava dušóno

Remove tab

Synonyms

pačúnnavom ad uvótre
páda mi húbovo na dušóso
alóknava mi na dušóno

Figure 6: Synonyms of (16c).

Other Relations

Relation ↑↓	Expression
Has_inchoative	alóknavá mi dušása
Verb Alternation	alóknava mi na dušóno
Opposite	izzéde mi dušóso

Figure 7: The Has_inchoative relation defined on (16a).

7 The future

Pomak-IDION is a unique resource of an endangered living language. It belongs to the set of Pomak resources developed in the framework of the project Philotis.

Pomak-IDION offers material and motivation for a future thorough study of Pomak VMWEs, e.g., studies on the role of the triple enclitic deictic article in idioms, the syntactic flexibility properties of VMWEs, verb alternation phenomena, language contact phenomena observed with LVCs and idioms and studies on the semantics of idiomatic Pomak.

Enriching IDION, and Pomak-IDION, with the inflectional paradigms of the VMWEs is among our future plans. This presupposes the encoding of the idiosyncratic constraints that hold for a number of VMWEs, other than constraints on the lexicalised parts: for instance, a VMWE may never appear in the future tense or the 1st person but may fully inflect for all the other tenses and persons. Such constraints are not expressed by the UD representation of the lemma form and are only partially covered by the corpus material; at the moment, encoders keep notes in IDION describing these properties of the VMWEs.

The development of the Pomak edition of IDION has shown that it can accommodate detailed information on VMWEs of different languages. In the future, cross-edition relations between VMWEs may be added to IDION. So far, each edition has been independent of the others; as a result, switching between the respective editions is required in order to see two equivalent expressions in two different editions. The implementation of cross-edition relations is an interesting documentation capability that will facilitate comparative studies on idiomaticity and other linguistic activities such as teaching and translation.

Abbreviations

GE	Modern Greek equivalent
LVC	Light verb construction
NLP	Natural Language Processing
K&K alphabet	Alphabet by R. Karahođa and P. G. Krimpas
PoS	Part of speech
UD	Universal Dependencies
VMWE	verbal multiword expression

Acknowledgements

We acknowledge full support of this work by the project “PHILOTIS: State-of-the-art technologies for the recording, analysis and documentation of living languages” (MIS 5047429), which is implemented under the “Action for the Support of Regional Excellence”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Appendix A Pomak LVCs

- (17) a. **fátom** **nazára**
 catch.VERB.1SG evil eye.NOUN
 ‘I am jinxed’. Verb: *nazarjásavom* ‘I am affected by evil eye’
- b. **stánavom** **budalá**
 become.VERB.1SG mad.ADJ
 ‘I go crazy’. Verb: *pabudalævom* ‘I go crazy’
- c. **stánavom** **dløg**
 become.VERB.1SG tall.ADJ
 ‘I grow tall’. Verb: *izdlógnavom* ‘I grow tall’
- d. **stánavom** **gulæm**
 become.VERB.1SG big.ADJ
 ‘I grow big’. Verb: *nagulæmávom* ‘I grow big’
- e. **stánavom** **hazýr**
 become.VERB.1SG ready.ADJ
 ‘I get ready’. Verb: *hazyrladísavom so* ‘I get ready’
- f. **stánavom** **star**
 become.VERB.1SG old.ADJ
 ‘I grow old’. Verb: *sastarævom, stárem* ‘I grow old’
- g. **stánavom** **zengínin**
 become.VERB.1SG rich.ADJ
 ‘I become rich’. Verb: *zenginjásavom* ‘I become rich’
- h. **tavárem** **so** **græha**
 load .VERB.1SG myself.PRON sin.NOUN
 ‘I commit a sin’. Verb: *græhóvom* ‘I commit a sin’

Appendix B Pomak idioms

- (18) a. **adbávem** **hatýrane**
destroy.VERB.1SG favour.NOUN
'I refuse to satisfy somebody's wishes' GE: *χαλάω χατίρι*
- b. **atkáčem** **jazýkate**
sever.VERB.1SG tongue.the.NOUN
'I make someone stop talking' GE: *κόβω τη γλώσσα κάποιου*
- c. **atvárem** **ačise**
open.VERB.1SG eyes.the.NOUN
'I realize what is going on' GE: *ανοίγω τα μάτια μου*
- d. **fáta** **gi** **sas** **annóš**
catch.VERB.3SG them.PRON with.ADP once.ADV
'he is bright' GE: *τα πιάνει με την μία*
- e. **fórnem** **něko** **na** **pótene**
throw.VERB.1SG somebody in.ADP street.the.NOUN
'I kick out someone' GE: *πετάω κάποιον στον δρόμο*
- f. **glódom** **tavánase**
look.VERB.1SG ceiling.the.NOUN
'I am absent minded' GE: *κοιτάω το ταβάνι*
- g. **hránem** **zmíje** **faf** **skútase**
feed.VERB.1SG snake.NOUN in.ADP bosom.the.NOUN
'I befriend somebody who proves to be deceitful' GE: *τρέφω φίδι στον κόρφο μου*
- h. **izzéde** **mi** **dušóso**
eat off.VERB.3SG.PST to/of-me.PRON soul.NOUN
'it has distressed me' GE: *μου έφαγε την ψυχή*
- i. **je** **mi** **so** **katá** **vólek**
be.AUX.1SG to/of-me.PRON REFL like.ADV wolf.NOUN
'I am starving' GE: *πεινάω σα λύκος*
- j. **na** **móžom** **da** **zgybem** **nagyse**
not.PART can.VERB.1SG that.ADP move.1SG leg.NOUN.PL
'I am exhausted, I am burnt out' GE: *δεν μπορώ να κουνήσω τα πόδια μου*

- k. **na pamína ad móse**
 not.PART go.3SG.VERB through.ADP my.the.SING.FEM.ACC
róky
 hand.NOUN.PL
 ‘it does not depend on me’ GE: *δεν περνάει από το χέρι μου*
- l. **na sésta so ad láfa**
 not.PART understand.3SG.VERB REFL from.ADP word.NOUN.PL
 ‘he is indifferent’ GE: *δεν καταλαβαίνει από λόγια*
- m. **na zaznáje mu so ušána**
 not.PART sweat.3SG.VERB to-me.PRON REFL ear.NOUN
 ‘I don’t give a damn’ GE: *δεν ιδρώνει το αυτί μου*
- n. **pádom na móko**
 fall.VERB.1SG on.ADP soft place.NOUN
 ‘I escape unpunished’ GE: *πέφτω στα μαλακά*
- o. **píjem bannómu karvtóno**
 drink.VERB.1SG somebody’s.PRON blood.NOUN
 ‘I drain somebody’s blood’ GE: *πίνω το αίμα κάποιου*
- p. **púkom ad játo**
 break.VERB.1SG from.ADP food.NOUN
 ‘I eat excessively’ GE: *σκάω από το φαγητό*
- q. **rábatem katá kúče**
 work.VERB.1SG like.ADV dog.NOUN
 ‘I work hard’ GE: *δουλεύω σαν σκύλος*
- r. **sédom sas svózany róky**
 sit.VERB.1SG with.ADP crossed.ADJ arms.NOUN
 ‘I do nothing, remain inactive’ GE: *κάθομαι με δεμένα χέρια*
- s. **videm bále déne**
 see.VERB.1SG white.ADJ day.NOUN
 ‘I get a break, I get ahead in life’ GE: *βλέπω άσπρη μέρα*
- t. **zímom go ad ustána mu**
 get.VERB.1SG it.PRON from.ADP mouth.NOUN his.PRON
 ‘I take the words out of somebody’s mouth’ GE: *το παίρνω από το στόμα του*
- u. **katá vadíca go naučem**
 like.ADV water.NOUN it.PRON learn.VERB.1SG
 ‘I learn something perfectly’ GE: *μαθαίνω νεράκι κάτι*

- v. **korv** **plújem** OR **kyrv** **hráčem**
blood.NOUN spit.VERB.1SG OR blood.NOUN spit.VERB.1SG
'I work hard to succeed' GE: *φτύνω αίμα*
- w. **máhnavot so** **káto** **dve** **kápky** **vódo**
look alike.3PL.VERB like.ADP two.NUM drop.NOUN.PL water.NOUN
'they are like peas in a pod' GE: *μοιάζουν σα δυο σταγόνες νερό*

References

- Adamou, Evangelia & Davide Fanciullo. 2018. Why Pomak will not be the next Slavic literary language. In D. Stern, M. Nomachi & B. Belić (eds.), *Linguistic regionalism in Eastern Europe and beyond: Minority, regional and literary microlanguages*, 40–65. Berlin: Peter Lang. <https://halshs.archives-ouvertes.fr/halshs-02105739>.
- Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 2nd edn., 267–292. Boca Raton, FL: CRC Press.
- Brenzinger, Matthias, Akira Yamamoto, Noriko Aikawa, Dmitri Koundioubá, Anahit Minasyan, Arienne Dwyer, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Osamu Sakiyama, María E. Villalón, Akira Y. Yamamoto & Ofelia Zepeda. 2003. *Language vitality and endangerment*. Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages.
- Constantinides, Nicolaos Th. 2007. *Η πομακική πολιτισμική μονάδα στην ελληνική Θράκη από άποψη Παρευξείνιων Σπουδών: Σύντομη ιστορική επισκόπηση, γλώσσα, ταυτότητες*. Democritus University of Thrace. (MA thesis).
- Constantinides, Nicolaos Th. 2020. Συγκλίσεις και αποκλίσεις στην Πομακική της ελληνικής Θράκης αφορούσες τα πεδία της αοριστίας, της οριστικότητας και του τριμερούς προσδιορισμού υπό το πρίσμα μιας σύνθετης λαογραφικής θέωρησης. ('Convergences and divergences in the Pomak of Greek Thrace concerning the fields of indeterminacy, finality and tripartite determination in the light of a complex folklore view'). *Mare Ponticum* 8(1). 56–76.
- de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

- Gantar, Polona, Lut Colman, Carla Parra Escartín & Héctor Martínez Alonso. 2018. Multiword expressions: Between lexicography and NLP. *International Journal of Lexicography* 32(2). 138–162. DOI: 10.1093/ijl/ecy012.
- Giouli, Voula, Vera Pilitsidou & Hephestion Christopoulos. 2024. A FrameNet approach to deep semantics for MWEs. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 147–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998639.
- Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1/2). 23–39. DOI: 10.1007/s10579-009-9094-z.
- Al-Haj, Hassan, Alon Itai & Shuly Wintner. 2013. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography* 27. 130–170. DOI: 10.1093/ijl/ect036.
- Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity* (Studies in Language Companion Series 23), 87–120. Amsterdam, Philadelphia: John Benjamins. DOI: 10.1075/slcs.23.05has.
- Hüllen, Werner. 2004. *A history of Roget's "Thesaurus"*. New York: Oxford University Press.
- Jespersen, Otto. 1965. *A Modern English grammar on historical principles*, vol. 6: Morphology. London: Allen & Unwin.
- Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali & Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 6282–6293. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.560. <https://aclanthology.org/2020.acl-main.560>.
- Karahóga, Ritván, Panagiotis G. Krimpas, Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karamatskos, Vasileios Sevetlidis, Nikolaos Constantinides, Nikolaos Kokkas, George Pavlidis & Stella Markantonatou. 2022. Morphologically annotated corpora of Pomak. In *Proceedings of the fifth workshop on the use of computational methods in the study of endangered languages*, 179–186. Dublin. DOI: 10.18653/v1/2022.computel-1.22. <https://aclanthology.org/2022.computel-1.22>.
- Karahóga, Sebajdín. 2017. *Μεταφράσεις ελληνικής και αγγλικής ποίησης στην πομακική γλώσσα*. Ξάνθη: Πολιτιστικός Σύλλογος Πομάκων Ξάνθης.
- Kokkas, Nikolaos. 2004. *Uchem so Pomátsko: Μαθήματα πομακικής γλώσσας*, vol. A. Ξάνθη: Πολιτιστικό Αναπτυξιακό Κέντρο Θράκης.

- Krimpas, Panagiotis G. 2020. Η γλώσσα και η καταγωγή των Πομάκων υπό το φως της Βαλκανικής Ζώνης Γλωσσικής Επαφής. In Manolis Varvounis, Antonis Bartsiakos & Nadia Macha-Bizoumi (eds.), *Οι Πομάκοι της Θράκης: πολυεπιστημονικές και διεπιστημονικές προσεγγίσεις* (Μελέτες Λαογραφίας και Κοινωνικής Ανθρωπολογίας 7), 167–204.
- Krimpas, Panagiotis G. 2022. Ευρωγλωσσολογία, νεοελληνική γλώσσα και ευρωπαϊκή ολοκλήρωση. In Zoe Gavriilidou, Nikolaos Mathioudakis, Maria Mitsiaki & Asimakis Fliatouras (eds.), *Γλωσσανθοί: Μελέτες αφιερωμένες στην Πηνελόπη Καμπάκη-Βουγιουκλή*, 153–169. Athens: Herodotus, Democritus University of Thrace.
- Laporte, Éric. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 143–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.1182597.
- Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.
- Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.
- Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann & Johanna Monti. 2016. PARSEME survey on MWE resources. In *10th international conference on Language Resources and Evaluation (LREC 2016)*, 2299–2306. Portorož. <https://hal.science/hal-01316351>.
- Lyons, John. 1977. *Semantics*, vol. 1. Cambridge University Press. DOI: 10.1017/CBO9781139165693.
- Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019) at ACL 2019*, 130–134. Florence. DOI: 10.18653/v1/W19-5115.
- Markantonatou, Stella, Nicolaos Th. Constantinides, Vivian Stamou, Vasileios Arampatzakis, Panagiotis G. Krimpas & George Pavlidis. 2023. Methodological issues regarding the semi-automatic UD treebank creation of under-resourced

- languages: The case of Pomak. In *Proceedings of the sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, 27–35. Washington, D.C. <https://aclanthology.org/2023.udw-1.4>.
- Ní Loingsigh, Katie & Brian Ó Raghallaigh. 2016. Starting from scratch: The creation of an Irish-language idiom database. In Tinatin Margalitadze & George Meladze (eds.), *Proceedings of the 17th EURALEX International Congress*, 726–734. Tbilisi: Tbilisi State University.
- Nielsen, Jakob & Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'90)*, 249–256. DOI: 10.1145/97243.97281.
- Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginia Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.
- Papadimitriou, Panayotis. 2008. *Τα Πομάκια: Συγχρονική περιγραφή μιας νότιας τοπικής ποικιλίας της αναλυτικής σλαβικής από τη Μύκη του Ν. Ξάνθης*. Thessaloniki: Kyriakides Bros.
- Papadimitriou, Panayotis. 2013. *Λαλιές Πομάκων της ελληνικής Ροδόπης: Περιφερειακή Αναλυτική Σλαβική και μουσουλμάνοι ομιλητές στη Νοτιοανατολική Ευρώπη*. Thessaloniki: Institute for Balkan Studies.
- Piirainen, Elisabeth. 2005. Europeanism, internationalism or something else? Proposal for a cross-linguistic and cross-cultural research project on widespread idioms in Europe and beyond. *HERMES: Journal of Language and Communication in Business* 18(35). 45–75. DOI: 10.7146/hjlc.v18i35.25816. <https://tidsskrift.dk/her/article/view/25816>.
- Sailer, Manfred & Stella Markantonatou. 2016. Affectees in MWEs: German and Modern Greek. In *Posters from the PARSEME 6th general meeting, 7–8 April 2016, Struga, North Macedonia*. <https://typo.uni-konstanz.de/parseme/images/Meeting/2016-04-07-Struga-meeting/WG1-MARKANTONATOU-SAILER-poster-1.pdf>.
- Sandry, Susan. 2013. *Phonology and morphology of Paševik Pomak with notes on the verb and fundamentals of syntax*. University College London, School of Slavonic & East European Studies. (MA thesis).
- Savary, Agata. 2009. Multiflex: A multilingual finite-state tool for multi-word units. In Sebastian Maneth (ed.), *Implementation and application of automata*, 237–240. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-02979-0_27.

- Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejcek, Fabienne Cap, Slavomir Ceplo, Silvio Ricardo Cordeiro, Gulsen Eryigit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartin, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.14715.
- Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the joint workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91. Florence. DOI: 10.18653/v1/W19-5110.
- Skoumalová, Hana, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondříčka & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631.
- Theocharides, Petros. 1996a. *Γραμματική της Πομακικής Γλώσσας*. Thessaloniki: Egiros.
- Theocharides, Petros. 1996b. *Πομακο-Ελληνικό Λεξικό/Πομάχτσκου-Ουρούμτσκου Λεκσικό*. Thessaloniki: Egiros.
- Villavicencio, Aline, Timothy Baldwin & Benjamin Waldron. 2004. A multilingual database of idioms. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC'04)*, 1127–1130. Lisbon. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/760.pdf>.
- Vondříčka, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112. 83–101.

Chapter 3

A uniform multilingual approach to the description of multiword expressions

👤 Svetlozara Leseva^a, 👤 Verginica Barbu Mititelu^b, 👤 Ivelina Stoyanova^a & 👤 Mihaela Cristescu^c

^aDepartment of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences ^bResearch Institute for Artificial Intelligence, Romanian Academy ^cFaculty of Letters, University of Bucharest

In this chapter we describe a linked bilingual (Bulgarian and Romanian) computational lexicon of multiword expressions, a new resource which encompasses lexical, morphological, semantic and stylistic information, in an independent, though unified way. The lexicon is a bilingual lexicographic resource, originating in the wordnets for the two languages, and is made up of self-contained monolingual lexicons of multiword expressions, which may be expanded to cover other levels and features of linguistic description, as well as other languages.

1 Introduction and main objectives

Along with the efforts in the domain of traditional lexicography, various developments towards the compilation of lexicons of multiword expressions (MWEs) for the needs of computational lexicography and computational linguistics have also been undertaken. As emphasised in a position paper (Savary et al. 2019) that emerged from the PARSEME¹ initiative (Savary et al. 2015), devising syntactic

¹PARSEME was a COST Action (2013–2017) focusing on parsing and MWEs. Some of its major results were the creation of annotation guidelines for verbal MWEs for more than 20 languages from various language families, a multilingual journalistic corpus annotated according to these guidelines made publicly available and a series of shared tasks on the identification of MWEs in texts, in which the previously mentioned corpus was used for training and testing the participating systems. See <https://typo.uni-konstanz.de/parseme/>.



MWE lexicons was recognised as a prerequisite for advancing research in MWE identification and other MWE-related tasks.

We propose an electronic bilingual MWE lexicon that comprises morphological (inflectional and derivational alike), syntactic (including word order) and semantic description in an independent, though unified, way. We build upon the one proposed by Leseva et al. (2020), itself inspired by the MWE description in Koeva et al. (2016). Our goal is to create a linked bilingual lexicographic resource consisting of self-contained monolingual lexicons of MWEs that may be expanded to other levels of linguistic description and to other languages.

Our work has the following main contributions: (i) an overview of several approaches for the description of MWEs with interest in language-independent, cross-lingual, bilingual, and/or multilingual representation, and especially in the features used in the MWE description – see §2. §3 briefly describes the wordnets² for the two languages in focus and their characteristics that allow for the creation of the linked lexicon presented here, along with the compilation of the datasets of verbal multiword expressions (henceforth VMWEs) involved in the linguistic analysis. The features previously mentioned serve as a starting point in designing the structure of the MWE lexicons for Bulgarian and Romanian, linked into one resource, described in this work; (ii) the presentation of a uniform framework for the construction of a linked resource consisting of two MWE lexicons (for Bulgarian and Romanian) that takes into consideration the advantages and challenges posed by the existing approaches and practices – see §4. This is a first step in the creation of a multilingual resource for the lexicographic description of MWEs, both in structural and semantic perspectives; (iii) the exploration of the lexicographic representation of MWEs in the context of aligned general lexical, semantic and morpho-syntactic resources not exclusively compiled for MWEs, this step being an important prerequisite for various Natural Language Processing (NLP) applications – see §5. We show that a uniform description of MWEs is possible for two languages from different families, highlighting language similarities, but also ensuring the mechanisms that allow for the description of language specificities.

²We write *wordnet* when referring to a “lexical knowledge base for a given language, modeled after the principles of Princeton WordNet” (see http://www.dblab.upatras.gr/balkanet/journal/20_BalkaNetGlossary.pdf). We write *Wordnet* when referring to a particular such resource, here the Bulgarian Wordnet and the Romanian Wordnet; the form *WordNet* is used only with reference to the trademarked Princeton WordNet (see <https://wordnet.princeton.edu/>).

2 Advances in computational lexicography with a recourse to MWEs

Most of the times, MWEs are recorded in general language dictionaries, where they are usually only semantically described, i.e., their meaning is explained. Large computational lexical resources also make provisions to incorporate MWEs (Chiarcos et al. 2024 [this volume]). Even valence dictionaries focused on the general language can contain descriptions of MWEs: see Walenty (Przepiórkowski et al. 2014b), which was extended to accommodate properties of MWEs (Przepiórkowski et al. 2014a).

However, dedicated lexicons do exist for MWEs in some languages and various grammatical formalisms were adopted in their description: the Lexicon-Grammar framework (Gross 1975, 1982), which spurred substantial advances in the formal linguistic description, including the treatment of MWEs, was more recently used in the description of Italian MWEs (Vietri 2014b, Monti 2014); Lexical-Functional Grammar (LFG, Bresnan 1978, Dalrymple 2023) was applied in the development of a Norwegian MWE resource (Dyvik et al. 2019); Head-driven Phrase Structure Grammar (HPSG, Pollard & Sag 1987, 1994, Müller et al. 2021) was adopted in the LinGO project³ for the creation of a lexicon including both simplex entries and MWEs (Villavicencio et al. 2004b); Frame Semantics was used to provide shallow semantic representation of multiword predicates (Giouli et al. 2024 [this volume]); Meaning-Text Theory (Mel'čuk 1981) was employed in Mel'čuk's (2006) Explanatory Combinatorial Dictionary, while the work by Schafroth (2015) offers a learner-centered description of Italian idioms based on the theoretical principles of Construction Grammar (Fried & Östman 2004).

Most MWE lexicons are monolingual resources (Fellbaum & Geyken 2005, Grégoire 2007, Odijk 2013, Shudo et al. 2011, Villavicencio et al. 2004b, Vietri 2014b, Schafroth 2015, Mel'čuk 2006, Markantonatou et al. 2019, Skoumalová et al. (2024 [this volume])). Others boast multilinguality as an important feature. However, multilingual support is ensured in different ways in different projects. Villavicencio et al. (2004a) report on MWEs in a source language that are manually given their equivalents in a target language, thus ensuring semantic equivalence between MWEs in the two languages, while the lexical and syntactic equivalences have to be decided upon by the user. Konbitzul⁴ (Iñurrieta et al. 2018) is a bilingual Spanish-Basque verb-noun lexicon of MWEs. Besides containing MWE equivalents in the two languages, it also offers morphosyntactic information about the MWEs in both languages, which is introduced either manually

³<https://www-csli.stanford.edu/groups/lingo-project>

⁴<http://ixa2.si.ehu.es/konbitzul/>

or semi-automatically. The Genoese-Italian phraseological dictionary⁵ describes Genoese MWEs, including their Italian equivalent(s) (Autelli 2020).

Some of the discussed MWE initiatives supply translation equivalents to the described units in other languages (either MWEs, if available, or free phrases) (Markantonatou et al. 2019, 2020, 2024). This feature is especially useful for dictionaries of less-spoken languages where the use of English as a metalanguage increases the usability and understandability of the resource.

Some of the projects developing MWE resources focused on harvesting them from corpora, providing consistent representation of the MWE system within a language, as well their extensive description at various linguistic levels.

Harvesting of MWEs from corpora was done (i) automatically, either from corpora annotated with MWEs (Grégoire 2007) or from corpora lacking such annotation (Fellbaum & Geyken 2005, Odijk 2013); or (ii) manually (Dyvik et al. 2019, Shudo et al. 2011, Odijk et al. 2024).

Given the characteristics of MWEs (e.g., discontinuity, inflection of components, word order variation, etc.), the automatic analysis of corpora is prone to errors, hence it is usually followed by a manual inspection and selection of MWEs. Automatic identification of MWEs in corpora benefits from the morphosyntactic annotation and lemmatisation of the texts (Odijk 2013). Some authors combine the extraction of MWEs from corpora with selecting MWEs from available idiom or general-purpose dictionaries or lists. In such cases, examples from corpora and/or the web serve to supplement the dictionaries with new entries, to confirm and exemplify the uses and various phenomena concerning MWEs (Hnátková et al. 2019, Markantonatou et al. 2019, 2020, Skoumalová et al. 2024).

Describing the system of MWEs within a language concerns the paradigmatic aspect of MWEs, a topic that is more rarely touched upon in the dedicated literature. Grégoire (2007) discusses the organisation of Dutch MWEs in classes (called “equivalence classes”) according to syntactic characteristics, the inner structure of MWEs and the possibility for them to have modifiers; Villavicencio et al. (2004b) use “meta-types” to organise the MWEs in classes and to map “the semantic relations between the elements of the MWE into the appropriate grammar dependent features” (Villavicencio et al. 2004b).

With respect to the way in which MWEs are described in lexicographic resources, two trends were dominant in the literature. In one of them, all MWEs are entries in a lexicon: their description is made either by specifying a class to which they belong (Grégoire 2007) or by enumerating their characteristics, with

⁵<https://romanistik-gephras.uibk.ac.at>

special focus on idiosyncrasies (Gross 1996, Shudo et al. 2011, Al-Haj et al. 2013, Markantonatou et al. 2019, 2020).

In a different approach, Villavicencio et al. (2004b) propose a description of MWEs adjusted to their decomposable or non-decomposable types. Thus, fixed (i.e., non-decomposable) MWEs should be treated as simplex entries: their orthography, syntactic and semantic type as well as morphological inflection of components are specified. Flexible or decomposable expressions are also lexical entries encoded in three stages: (i) their components are registered as idiomatic entries associated with the non-idiomatic entries from which they inherit their grammatical characteristics; (ii) over-generation is avoided by defining the context of use for these idiomatic entries: for each MWE the components are listed, along with their obligatory or optional status; (iii) MWEs are assigned to a meta-type.

Similarly, Al-Haj et al. (2013) include MWEs as entries in their lexicon, alongside entries of simple words. Each component of a MWE contains a pointer to the corresponding simple entry in the lexicon. In a way similar to Villavicencio et al. (2004b), they propose adding fossil words⁶ as entries, which are not assigned a part of speech, but are marked as “fossil”, which is an indication of their occurrence only as components of MWEs.

Alternatively, in the *Explanatory combinatorial dictionary* (Mel’čuk 2006) different types of MWEs are treated differently: idioms and quasi-idioms are allotted separate entries (also cross-referenced with their components’ entries) with their own fully-fledged description, whereas the so-called semi-phasemes are described in the entry of their base, which, in the case of light verb constructions (LVCs) (a type of semi-phasemes), is most often a noun serving as the semantic head of the expression. The combinatorial properties of semi-phasemes are represented lexicographically by means of a special lexical function. Equivalent meanings formed on different support verbs are listed together.

Given that no standard was defined for it (yet), an important aspect of the linguistic description of MWEs is that it should not be framework-specific and should allow for its reuse by any system (Odijk 2013). There is agreement among researchers that MWEs must be explicitly marked as such in lexicons (Fellbaum & Geyken 2005, Mel’čuk 2006, Al-Haj et al. 2013, Dyvik et al. 2019, Hnátková et al. 2019, Markantonatou et al. 2019, 2020).

Taking as a point of departure the above mentioned lexicographic resources that focus on or include MWEs, below we summarise the levels of description we consider relevant for our work: lexical, derivational, morphological, syntactic,

⁶Fossil words are those that only occur in MWEs; they are also known as *cranberry* words.

semantic, contextual, stylistic.⁷ A detailed description of the complex multilevel representation of a broad range of MWEs and MWE types in Czech (another morphologically rich language), which shares many commonalities with the approach adopted herein is presented in Skoumalová et al. (2024 [this volume]). A different, though not contradicting approach to a rich multilayered description for Bulgarian MWEs is adopted in Osenova & Simov (2024 [this volume]). We defer the discussion as to which levels of description are implemented (and how) in the proposed Bulgarian-Romanian VMWE lexicon to §4, where we also provide an explanation for favouring a particular decision or approach over another.

2.1 Lexical level

The lexical level contains information about:

- the list of lexemes that can substitute components in the multiword expressions (Villavicencio et al. 2004b, Grégoire 2007, Przepiórkowski et al. 2014a, Hnátková et al. 2019, Markantonatou et al. 2019, 2020, Skoumalová et al. 2024). The variations may be handled uniformly regardless of the status of the component affected (i.e., as alternative realisation within the same citation form) or differently, according to certain criteria, e.g., whether the verbal head or an invariable component is concerned, cf. the treatment by Markantonatou et al. (2019, 2020);
- cross-references from the dictionary entries of each of the components of the MWEs (except for function words) to the entry/ies of the MWEs in which they occur (Villavicencio et al. 2004b, Mel'čuk 2006).⁸

2.2 Derivational information

Expressions that are derivationally related to the MWEs, e.g., nominal expressions derived from VMWEs (Mel'čuk 2006, Hnátková et al. 2019, Monti 2014), are recorded in the dictionary, thus providing links to other parts of the language's lexicon, including MWEs and one-word compounds.

⁷For a discussion of lexical encoding formats for MWEs that can be used in NLP systems, see Lichte et al. (2019).

⁸In Mel'čuk (2006) it is not clear if all lexical entries of a MWE component contain references to the respective MWE or only that which reflects the meaning it has in the MWE, although the author admits the semantic non-compositionality of some idioms.

2.3 Morphological description

The following information pertain to this level:

- lemma (i.e., canonical) form of all the components (Dyvik et al. 2019, Grégoire 2007, Odijk 2013, Odijk et al. 2024, Osenova & Simov 2024, Skoumalová et al. 2024);
- restrictions on the inflection of components that can help automatically generate all the possible forms of the MWE (Grégoire 2007, Al-Haj et al. 2013, Markantonatou et al. 2019, 2020, 2024, Osenova & Simov 2024, Skoumalová et al. 2024).

2.4 Syntactic level

This level contains the following information:

- syntactic category of the expression (e.g., nominal, verbal, adjectival, etc.) (Shudo et al. 2011, Al-Haj et al. 2013, Dyvik et al. 2019, Markantonatou et al. 2019), sometimes referred to indirectly, by means of reference to the class to which the MWE belongs (Grégoire 2007, Odijk 2013);
- internal syntactic structure of the expression (Dyvik et al. 2019, Grégoire 2007, Hnátková et al. 2019, Markantonatou et al. 2019, 2020, Shudo et al. 2011, Przepiórkowski et al. 2014a, Villavicencio et al. 2004b, Mel'čuk 2006) represented in terms of one of various theoretical frameworks: dependency structures (Hnátková et al. 2019, Odijk 2013, Villavicencio et al. 2004b, Markantonatou et al. 2024, Osenova & Simov 2024, Skoumalová et al. 2024), Lexicon-Grammar (Gross 1982), HPSG (Villavicencio et al. 2004b), LFG (Dyvik et al. 2019), constituent structures (Skoumalová et al. 2024 [this volume]), among others;
- possible modifiers of components (Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, Grégoire 2007, Shudo et al. 2011, Al-Haj et al. 2013, Markantonatou et al. 2024, Osenova & Simov 2024, Skoumalová et al. 2024);
- clear indication of the optional and obligatory components (Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, Villavicencio et al. 2004b, Markantonatou et al. 2024, Skoumalová et al. 2024);
- word order of the components with respect to each other (Al-Haj et al. 2013, Markantonatou et al. 2019, 2020, 2024) or marking of specific or anomalous

word order Markantonatou et al. (2024 [this volume]), Skoumalová et al. (2024 [this volume]), see also the approach adopted below;

- valency information about the MWE which determines its realisation in text (Giouli et al. 2024 [this volume]), (Osenova & Simov 2024 [this volume]), (Skoumalová et al. 2024 [this volume]);
- combinatorial possibilities of the expression extracted from corpora, such as possible subjects, complements, pre- or post-modifiers, etc. (Odičk 2013, Mel'čuk 2006), sometimes with their frequency (Odičk 2013, Odičk et al. 2024);
- other syntactic variations such as passivisation, causative-inchoative alternations, long-distance dependencies, alternative forms of the MWE (Dyvik et al. 2019, Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, Vietri 2014a, Markantonatou et al. 2024, Skoumalová et al. 2024), but only when they violate the rules of the grammar (Mel'čuk 2006).

2.5 Semantic description

The information contained at this level consists of:

- a paraphrase, a definition or an explanation of the meaning of the MWEs (Villavicencio et al. 2004b, Markantonatou et al. 2019, 2020, Mel'čuk 2006, Osenova & Simov 2024, Markantonatou et al. 2024, Skoumalová et al. 2024);
- relations to other idioms, such as synonymy (Autelli 2020, Osenova & Simov 2024, Markantonatou et al. 2024), antonymy (Fellbaum & Geyken 2005, Markantonatou et al. 2019, 2020, 2024), hypernymy and hyponymy (Fellbaum & Geyken 2005), as well as other relations that serve to define a network of VMWEs expressing a concept (Markantonatou et al. 2019, 2020): causative-inchoative or stative relations, verb alternations, lexical variants, etc., making it possible to group MWEs in synonym sets (Markantonatou et al. 2019, 2020, 2024);
- semantic domain (Fellbaum & Geyken 2005, Monti 2014), by means of cross-references to other entries in the dictionary having the same or related meaning (Mel'čuk 2006).

2.6 Contextual information

This level contains information such as:

- examples of sentences (extracted from corpora) containing the respective MWE (Grégoire 2007, Markantonatou et al. 2019, 2020, Odijk 2013, Autelli 2020, Osenova & Simov 2024, Markantonatou et al. 2024, Skoumalová et al. 2024). When the MWEs in the lexicon originate from corpora, the information extracted from the corpus (such as context of occurrence, frequency, etc.) is kept track of by a reference from the lexicon entry to the file storing the respective information (Grégoire 2007);
- contextual restrictions on the occurrences of MWEs, such as co-occurrence with specific syntactic phrases (Shudo et al. 2011) or with semantically specific adverbs or other external modifiers (Fellbaum & Geyken 2005);
- frequency of occurrence of MWEs in corpora (Odijk 2013).

2.7 Stylistic information

The label “stylistic” encompasses all kinds of information about the style or language register in which a MWE is typically used, such as “ironic”, “disparaging”, “humorous” (Fellbaum & Geyken 2005); “formal”, “colloquial”, “offensive” (Markantonatou et al. 2019, 2020); “vulgar”, “negative connotation”, “disused” (Autelli 2020), or other similar descriptions (Skoumalová et al. 2024 [this volume]).

2.8 Other information

Besides the linguistic types of information already mentioned, some lexicons also include the following:

- diachronic information: changes in the form and meaning of the VMWEs over time (Fellbaum & Geyken 2005);
- translation into other languages such as English (Al-Haj et al. 2013, Markantonatou et al. 2019, 2020) and French (Markantonatou et al. 2019, 2020, 2024);
- the emphatic function of MWEs (Fotopoulou et al. 2014, Markantonatou et al. 2019, 2020).

The overview of the types of linguistic information encoded about MWEs shows that the lexicons referenced above contain relevant descriptions and partially overlapping types of information, distributed over several linguistic levels. One of our aims when developing the linked Bulgarian-Romanian bilingual lexicon of MWEs was to provide a consistent and uniform framework for the representation of MWEs that would take into account the various levels of linguistic description and the approaches to tackle them in line with the findings of the theoretical analysis as well as the specific requirements of the bilingual (and by extension – multilingual) representation of data.

3 Compilation of a Bulgarian-Romanian MWE lexicon

We first describe the lexical resources that the lexicon is derived from, i.e., the Bulgarian and Romanian wordnets. We then present the different levels of linguistic description in comparison with other frameworks and initiatives.

3.1 BulNet and RoWN: Sources of MWEs for the lexicon

A wordnet is a semantic network: its nodes are represented by synonym sets (synsets), which contain one or more linguistic items (called “literals”) that lexicalise a concept; literals may be single words or multiword combinations alike.⁹ The edges connecting the nodes are semantic relations that hold between a pair of synsets. Only words belonging to content parts of speech are usually represented in such language resources: nouns and verbs have a hierarchical organisation, descriptive adjectives are organised in clusters created around a pair of antonymic adjectives, relational adjectives and adverbs have no organisation. The first such network, Princeton WordNet (WordNet, Miller 1995), was developed for English; wordnets for other languages have been subsequently developed,¹⁰ most of which are aligned with WordNet, i.e. the synsets in different wordnets with equivalent meanings are mapped to each other.

The development of the Bulgarian Wordnet (BulNet, Koeva 2010) and the Romanian Wordnet (RoWN, Tufiş & Barbu Mititelu 2014) started in the BalkaNet project (Tufiş et al. 2004), which had as one of its objectives the implementation of a set of synsets common to all languages in the project. The construction of the two wordnets adopted the “expand” approach, which involves translation of the

⁹For a discussion on the representation of figurative language, proverbs and idioms in WordNet, see Fellbaum (1998a).

¹⁰For a list of existing wordnets in the world, see <http://globalwordnet.org/resources/wordnets-in-the-world/>.

literals in the English synsets and automatic transfer (and possibly revision) of the semantic relations from WordNet (Fellbaum 1998b) to BulNet and RoWN. The content of the synsets and associated information (literals, gloss, usage examples, stylistic notes, etc.) were devised by native language experts, who consulted relevant monolingual and bilingual dictionaries. These decisions and work methods led to the creation of wordnets aligned to WordNet and thereby to each other (via WordNet),¹¹ on the other. Figure 1 shows the interlinking among the wordnets, in which the English, Romanian and Bulgarian synsets contain verbal idioms: (bg) *давам най-доброто от себе си* *cu davam nay-dobroto ot sebe si* (lit. ‘give the best of oneself’), *давам всичко от себе си* *cu davam vsichko ot sebe si* (lit. ‘give all of oneself’) – (ro) *da totul* (lit. ‘give all’), *da ce e mai bun* (lit. ‘give the best’), *da tot ce e mai bun* (lit. ‘give all the best’).

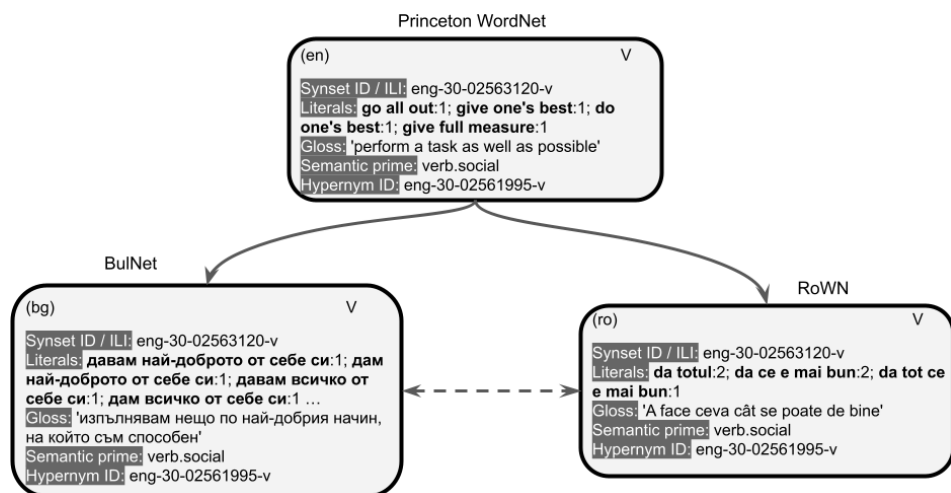


Figure 1: Interlinking wordnets.

After the end of BalkaNet, each team continued the development of the respective wordnet independently, with different interests in the conceptual coverage of their resources. The development of the wordnets for Bulgarian and Romanian (as well as for any other language constructing a wordnet using the expand approach) is naturally biased towards English, as WordNet provided the original inventory of senses. While this fact was acknowledged, it was not considered a serious concern, as no resource could be absolutely unbiased, on the one hand, and because of the fact that concepts are shared by different languages, which made the alignment among wordnets possible, on the other. MWEs were not

¹¹They are also aligned to any wordnet that is aligned to WordNet.

a particular focus of the development of BulNet and RoWN; however, as they are treated on a par with single words, MWEs were included whenever relevant for a synset. The current versions of the two wordnets do not cover the lexical inventory of the languages thoroughly.¹²

3.2 Dataset construction

The features of the bilingual resource outlined in the following sections were described on the basis of linguistic analysis aiming at delineating the common linguistic characteristics and the differences between the two languages that need to be taken into account in such a lexicon. This analysis is based on 3,656 multitoken literal-to-literal pairs in corresponding synsets in BulNet and RoWN. These include VMWEs proper, as well as multitoken free phrases with purely compositional meaning. We filtered out the latter and were left with 2,705 VMWE-to-VMWE pairs. As the VMWEs under discussion are part of pairs of corresponding aligned synsets, they are treated as possible translation equivalents to each other, cf. the synset counterparts in (1), and are included in the constructed bilingual resource. As part of the VMWE bilingual lexicon, each VMWE is analysed and described on the morphological, syntactic, semantic, stylistic, connotational and derivational level individually. The linguistic information which is common to all the members of a synset, e.g. the gloss, is also assigned to each VMWE in the relevant synset, as each VMWE is a separate unit in the VMWE lexicon. In addition, all the VMWEs belonging to the same synset share the same synset ID and are thus identifiable as part of the synset. We did not implement any further linking beyond the alignment at the synset level, which was performed while the individual wordnets were being constructed.

The verbal multiword literals in BulNet and RoWN were manually annotated with the VMWE types from the PARSEME 1.2 guidelines:¹³ verbal idioms (VID), light verb constructions whose verb is semantically totally bleached (LVC.full), light verb constructions in which the verb adds a causative meaning to the noun (LVC.cause), inherently reflexive verbs (IRV), for both languages, while the category inherently adpositional verbs (IAV) was annotated only for Bulgarian (Barbu Mititelu et al. 2019b).

The compilation of the lexicon started with those synsets that are lexicalised by VMWEs of the same type in both wordnets: 192 VID examples, 44 LVC ones and 2,023 IRVs. IRVs are also of interest for comparative studies, but will be part

¹²We used Princeton WordNet – 3.0 aligned with Bulgarian and Romanian wordnets. BulNet consists of 85,954 synsets created by expert linguists (Koeva 2021), while RoWN contains 59,348 synsets (Tufiş & Barbu Mititelu 2014).

¹³<https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/>

of future work. Thus, the set of VMWEs currently included into the lexicon and subject to description is made up of 2,259 pairs of corresponding VMWEs.

The description of VMWE literals was performed independently for each of the two languages according to a common set of features and their possible values. IRVs have regular structure, word order and syntactic properties, so our work is focused only on VID and LVC cases, which pose a number of challenges for their description and the analysis of their properties.

As a result, we obtain a new resource, a self-contained bilingual MWE lexicon where each VMWE in each of the languages is described individually, but each VMWE is described by filling in the relevant fields in the predefined template of a language-independent lexicon entry. In the following section we delve into the types of information included in each dictionary entry and how these are handled in practical terms.

4 The content of a lexicon entry

Following one of the dominant trends in MWE lexicon crafting, we adopt the approach of encoding VMWEs explicitly as distinct entries instead of describing the rules of combining their components. This makes it possible to reflect and access in a straightforward way the morphosyntactic, syntactic, semantic and derivational information associated with a particular entity that may not be readily obtainable from the combination of its components. In (1), we illustrate three aligned synsets in WordNet, BulNet and RoWN.¹⁴ We notice that in the same synset there may be MWEs based on a different support verb (as in (2a) for Bulgarian) or a different semantic head (as in (2b) for Romanian).¹⁵

- (1) a. form:8; take form:1; take shape:1; spring:6 (en)
 Synset ID: eng-30-02623906-v
 Definition: ‘develop into a distinctive entity’
- b. образувам се:1, оформям се:2, оформя се:2, формирам се:1,
 obrazuvam se:1, oformyam se:2, oformya se:2, formiram se:1,
 приемам форма:1, приема форма:1, добивам форма:1, добия
 prieam forma:1, priema forma:1, dobivam forma:1, dobiya
 форма:1, кристализирам:1 (bg)
 forma:1, kristaliziram:1
- c. se forma:1; se contura:1; prinde contur:1; prinde formă:1 (ro)

¹⁴The synset ID and definition are rendered only for WordNet.

¹⁵For brevity, we do not give literal translations where they are similar or identical to the idiomatic translation.

- (2) a. приемам форма, добивам форма (bg)
priemam forma, dobivam forma
adopt shape, obtain shape
- b. prinde contur, prinde formă (ro)
catch outline, catch shape

While it is obvious that some literals are closer correspondences to each other in terms of structure and/or semantics – e.g. (bg) *formiram se* – (ro) *se forma* (‘form’) and (bg) *priemam forma* – (ro) *prinde formă* – (en) *take form* – we do not attempt to connect to each other such stricter correspondences found within the same pairs of synsets; instead, we take all literals on one side to be relevant translation equivalents of the literals on the other side, as translation choices may be guided by factors other than structural or semantic similarity.

In the following subsections we present the levels of description of VMWEs adopted in the resource presented. Given one of the organisation principles of WordNet, i.e., each synset stands for a concept and each word/expression can occur a number of times equal to its number of senses, it is clear that all information provided for a MWE pertains to one of its senses, in case it is polysemous.

4.1 Technical information

This level of description serves two main purposes: the unique identification of the VMWE lexicon entry within the dataset for one language, as well as pairing the VMWE entries across languages. For this, we employ wordnet indexing with additional identification elements which serve both to identify a VMWE as part of a particular synset (via synset ID) and to distinguish it from other VMWEs in the same synsets, or from identical VMWE literals in other synsets (via literal IDs, see (3)). For Bulgarian we also include a verb aspect identifier, which allows us to refer jointly or separately to aspectual pairs lexicalising the VMWEs – this is useful when comparing languages that differ with respect to the verb aspect or where the aspectual systems are organised differently.¹⁶ The identification system allows us to: (i) access all the synset-level linguistic information provided; (ii) make references to a particular VMWE uniquely, e.g., in the description of derivatives (e.g., (3b) as derived from (3a) and not from its aspectual counterpart *snema otpechatatsi*, literal ID: bg_2330, nor its synonym *vzemam otpechatatsi*, literal ID: bg_2327); (iii) extract translation equivalents of VMWEs from wordnets

¹⁶This feature is only relevant for Bulgarian. Romanian lacks a lexico-grammatical verb aspect (i.e. marked on separate lexemes) and aspectual distinctions are expressed by other means.

for different languages; (iv) use the rich relational structure of WordNet for the purposes of the semantic description of VMWEs.

- (3) a. снемам отпечатъци (bg)
snemam otpechatatsi
take fingerprints
Synset ID: eng-30-01748748-v, Literal ID: bg_2329, Aspect: IPFV
- b. снемане на отпечатъци (bg)
snemane na otpechatatsi
taking of fingerprints (the act of fingerprinting)
Synset ID: eng-30-00152338-n

4.2 Morphological description

4.2.1 Lemma of the VMWE

Savary (2008) considers two main approaches to lemma representation that have become dominant: (i) an abstract lemma, where a citation form that generates all the possible forms of the relevant single word is assigned to each component; (ii) a non-abstract lemma in which each of the components is represented by the form that is part of the relevant MWE, and the MWE lemma is associated with a formalised description of the grammatically possible combinations of forms of the MWE components, thus avoiding overgeneration. Even though the latter approach is linguistically more justified and was adopted by other authors (see §2.3 above), the former allows recognition and retrieval of MWEs from corpora where MWEs are not annotated, thus possibly being capable of recognising MWEs not included in lexicons. Still others (Fellbaum & Geyken 2005) determine MWE lemmas on the basis of the frequency of occurrence, maintaining multiple citation forms where two or more dominant forms are relatively equally distributed. Such an approach accounts for the fact that the non-abstract MWE lemmas are often not morphologically unmarked and that they may occur preferentially in particular forms but not in others.

We adopt a two-way approach by assigning each MWE both a non-abstract lemma and an abstract one. The function of the former is to represent the most neutral form in which the components occur in the language. It is this lemma that we consider in determining the inflection of the MWE components that reflects the actual morphological restrictions imposed on the forms that need to be described in the fields dedicated to morphosyntactic restrictions. Consider the following examples of non-abstract lemmas:

- (4) затварям си очите (bg)
 zatvaryam si ochite
 close self.CL eye.PL.DEF
 lit. ‘close one’s eyes’
 ‘turn a blind eye’
- (5) închide ochii (ro)
 close eye.PL.DEF
 lit. ‘close the eyes’
 ‘turn a blind eye’

In examples (4) and (5), the verbal head’s inflection is unrestricted, whereas the nominal complement is only found in its plural definite form. In addition, in Bulgarian the reflexive possessive pronoun is in its short (clitic) form (which is invariable). The description of the relevant restrictions in the dictionary prevents the overgeneration of non-existing forms.

For the automatic recognition of MWEs, we also encode an abstract lemma for each MWE (see examples (6) and (7) corresponding to the VMWEs in (4) and (5), respectively); this means representing the nominal complement in its citation form, i.e., singular indefinite for both languages, respectively, and, in Bulgarian, representing the reflexive possessive in its base form, i.e. masculine singular indefinite, which changes in terms of the number, gender and definiteness of the possessed entity.¹⁷

- (6) затварям свой око (bg)
 zatvaryam svoj oko
 close self.REFL.POSS.M.SG eye.SG.INDEF
 lit. ‘close one’s eye’
 ‘turn a blind eye’
- (7) închide ochi (ro)
 close eye.SG.INDEF
 lit. ‘close eye’
 ‘turn a blind eye’

The abstract lemma is invoked when a sequence of words corresponding to a MWE in the lexicon is recognised as such in a lemmatised corpus (i.e. where,

¹⁷The long form of the reflexive possessive pronoun does not denote person, and the categories it inflects for (gender, number, definiteness) are features not of the possessor but of the entity possessed.

most often, lemmas are assigned to single words); it is itself a sequence of forms that will not be found in the language in an idiomatic meaning or is completely impossible, as the abstract lemma in example (6) above: *zatvaryam svoy oko*.

The abstract lemma thus matches the lemmas assigned in the corpus and allows for each occurrence of the relevant MWE in the corpus to be associated with the dictionary entry and the information it contains.

The components of the MWE are numbered and identified with respect to their position in the lemma and the abstract lemma. In this way the morphological features, the restrictions on a component's paradigm, as well as the blocking of modifiers and external elements between particular components can be precisely defined.

4.3 Syntactic description

The syntactic variability of VMWEs is much greater than expected despite the traditional understanding about the relatively fixed nature of the structure and linearity of VMWEs. In particular, many (V)MWEs exhibit the regular syntactic behavior of free phrases, including the possibility of intervening external elements that modify a particular element of the VMWE or the entire expression/sentence, various semantic-syntactic transformations, alternative complement expression, long-distance dependencies, etc. That is why we chose to describe only the deviations from the regular syntactic behavior of the MWEs.

The syntactic description of the VMWEs in the lexicon is based on the Universal Dependencies¹⁸ (UD) framework (de Marneffe et al. 2021). The choice for this framework was natural, in order to ensure a consistent treatment of the VMWEs in the wordnets and in the Bulgarian (Savary et al. 2018) and Romanian (Barbu Mititelu et al. 2019a) corpora created (alongside those for other languages) within PARSEME, and annotated with the same types of VMWEs. These corpora were automatically syntactically annotated using UDPipe (Straka 2018), with the syntactic relations defined in UD (Savary et al. 2023).

There are several types of syntactic information recorded in our resource: the internal structure of VMWEs, their valence frames, word order restrictions on their components and the possibility of other words to occur within the expressions. They are all discussed in what follows.

¹⁸<https://universaldependencies.org/>

4.3.1 Internal syntactic structure

The syntactic annotation of the VMWEs in the two wordnets with UD relations was done manually, with the aim of describing the number of components within each VMWE and the syntactic relations between them. The representation of the VMWE structure follows this convention: the head of the expression (i.e., the verb) followed by the UD relations that the other components of the VMWEs establish with the head or with other components. In the description of the internal structure of VMWEs, the order of these relations reflects the linear order of the components in the expression. For example, the internal structure of the VMWE (en) *kick the bucket* is $V + [\text{det} + \text{obj}]$. The square brackets indicate that the determiner (det) and the direct object (obj) are not both attached to the verb, but only the obj, whereas the other depends on it.

Table 1 shows only some of the most frequent syntactic structures that have correspondences in the analysed VMWEs in Bulgarian and Romanian, but variants of these structures are omitted. For example, patterns such as $V + \text{obj}$ and $V + \text{case} + \text{obl}$ can have as variants $V + \text{obj} + \text{amod}$ and $V + [\text{case} + \text{obl} + \text{amod}]$ in Romanian and $V + [\text{amod} + \text{obj}]$ and $V + [\text{case} + \text{amod} + \text{obl}]$ in Bulgarian, $V + [\text{nummod} + \text{obj}]$ in both languages, where the word order variations arise from the structural differences in the two languages, i.e., in Romanian modifiers usually follow the nominal head, whereas in Bulgarian they precede it.¹⁹

We did include several parallel patterns. They are given a somewhat different analysis – i.e., we construe the possessive clitic in their structure as $\text{expl}:\text{poss}$ in Romanian and as det in Bulgarian. But, in fact, they correspond to each other and translate in the same way. Such an example is illustrated by the pattern $V + \text{expl}:\text{poss}/\text{det} + \text{obj}$. Leaving the linguistic discussion aside, we treat them as equivalent, thus aiming at pointing out the essential commonalities instead of the less important differences.

When correlating the PARSEME VMWEs types with their valence frames we notice the following. According to PARSEME guidelines, a characteristic of LVCs is the fact that they are made up of a verb and a noun, the latter determining the semantics of the expression. In Romanian and Bulgarian most expressions of the type LVC.full have the internal structure $V + \text{obj}$, consider the chess term (ro) *da şah* and its counterpart (bg) *davam şah*, both literally meaning ‘give check’ and translated as ‘place into check’, or $V + [\text{case} + \text{obl}]$ – (ro) *lua în serios* (lit. ‘take in serious’) ‘treat seriously’ and (bg) *stigam do sporazumenie* (lit. ‘reach to an agreement’) ‘come to an agreement’. The instances of Romanian LVC.cause

¹⁹Where such syntactic patterns are presented, we stick to a uniform way of encoding them, e.g., $V + [\text{obj} + \text{amod}]$, disregarding the differences between the two languages.

3 A uniform multilingual approach to the description of MWE

Table 1: Frequent syntactic structures within VMWEs in Bulgarian and Romanian.

Syntactic pattern	Romanian example	Bulgarian example
V+obj	<i>avea grijă</i> lit. 'have care' 'take care'	<i>imam grizha</i> lit. 'have care' 'take care'
V+ [case+obl]	<i>citi printre rânduri</i> lit. 'read among lines' 'read between the lines'	<i>cheta mezhdu redovete</i> lit. 'read between the lines' 'read between the lines'
V+ expl:poss/det +obj	<i>își ține gura</i> lit. 'keep one's mouth' 'shut one's mouth'	<i>zatvaryam si ustata</i> lit. 'close one's mouth' 'shut one's mouth'
V+obj+ [case+obl]	<i>arunca praf în ochi</i> lit. 'throw dust in eyes' 'throw dust in the eyes'	<i>hvarlyam prah v ochite</i> lit. 'throw dust in eyes' 'throw dust in the eyes'
V+ expl:poss/det + [case+obl]	<i>își ieși din fire</i> lit. 'escape from one's temper' 'flip one's lid'	<i>plyuya si na petite</i> lit. 'spit on one's heels' 'head for the hills'
V+ expl:poss/det +obj+advmod	<i>își lua cuvintele înapoi</i> lit. 'take back one's words' 'take back one's words'	<i>vzemam si dumite nazad</i> lit. 'take back one's words' 'take back one's words'
V+nsubj+ [case+advmod]	<i>lua gura pe dinainte</i> lit. 'the mouth takes on ahead' 'let the cat out of the bag'	–
V+advmod+ det+nsubj	–	<i>mnogo mi znae ustata</i> lit. 'my mouth knows a lot' 'have a big mouth'

display two types of internal structures, i.e., $V + \text{xcomp}$ and $V + [\text{case} + \text{obl}]$. The same structures are also found in Bulgarian, compare (ro) *face public* ‘make public’ and (bg) *pravva raven* ‘make equal’, as well as (ro) *pune în circulație* and (bg) *puskam v obrashtenie* ‘put into circulation’. In Bulgarian we also attested LVC.cause with the structure $V + \text{obj}$, e.g., (bg) *pravva upoyka* (lit. ‘administer anesthesia’) ‘put under, anesthesise’.

The internal structure of VIDs is more diverse, though, given that they can even be/contain clauses: e.g., (ro) *bate fierul cât e cald* ‘strike the iron while it is hot’. The syntactic structures attested in the data are based primarily on a verb-complement or verb-modifier pattern, while the subject and another complement or modifier are part of the VMWE’s valence frame. This fact is reflected in Table 1, which shows that only a few examples including a subject are found in the data, cf. the last two rows – (ro) *lua gura pe dinainte* and (bg) *mnogo mi znae ustata*, where the nouns *gura* and *ustata*, respectively, are the subject of the verb.²⁰

Besides the patterns in Table 1, the data also contains a number of structures that are less represented in the bilingual lexicon due to its size. In fact, many of them are variations of the ones described in the table, e.g., $V + [\text{case} + \text{advmod}]$ (bg) *izlizam na otkrito* (lit. ‘come out in open’) ‘come to light’ is a variant of $V + \text{advmod}$; the patterns involving an expletive reflexive (expl:pv), such as $V + \text{expl:pv} + [\text{case} + \text{obl}]$ (bg) *makna se po petite* (lit. ‘drag oneself on someone’s heels’) ‘tag along’, are variations of the respective models based on the pattern $V + \text{obj}$, as the expletive blocks the direct object (reflexive verbs are intransitive).

4.3.2 Morphosyntactic description

The morphosyntactic description deals with the morphological properties of the head and the dependent components of the VMWEs and the ways in which each of the components varies morphologically as part of the expression. The way morphological variation is treated depends on the extent of variation, the way the MWE lemma is defined, etc. (see §2). Regarding its variability, each component may be unrestricted (i.e., the MWE component displays the full simple word paradigm), restricted (the MWE component’s forms vary grammatically, but it is restricted with respect to one or more grammatical categories) or fixed (the MWE component does not vary morphologically).

We adopt the practice that lack of any morphosyntactic restrictions is the default value for each component and hence not marked, whereas restrictions or invariability are explicitly defined in the respective field of the MWE entry. For instance, in the following equivalent examples – (ro) *pune pe fugă* (lit. ‘put on

²⁰The empty cells show that the pattern is not attested in the data, though may well be possible in the language.

run'), (bg) *obrashtam v byagstvo* (lit. 'turn into flight') 'rout out, oust, cause to flee' – the MWEs consist of a verbal head and an oblique expressed by a noun introduced by a preposition (V + [case + obl]). The verb may be found in any form and is thus unrestricted, prepositions in both languages are invariable, while the noun is only found in its singular indefinite form.

In the analysed data, most often the verbal head's paradigm is unrestricted, with just a few exceptions, e.g., (ro) *lua gura pe dinainte* (lit. 'the mouth takes on ahead') 'let the cat out of the bag'. Examples of such exceptions are the cases where: (i) the nominal subject is part of the MWE and therefore the verbal head agrees with it; or (ii) the subject's referent cannot be a participant in the communication; or (iii) the verb is otherwise restricted as in weather expressions, where it can only be in the third person singular, e.g., (ro) *ploua cu găleata* (lit. 'rains with bucket') and (bg) *vali kato iz vedro* (lit. 'rains as if out of a bucket') 'rain buckets'.

We note that the most frequent restriction found in both languages is the singular indefinite form of the nominal dependent, followed by the singular definite form, etc. (Table 2). These restrictions are found across the most well-represented syntactic patterns – V + obj and V + [case + obl] as well as in more complex variations of these structures, e.g., V + [case + amod + obl]. – (bg) *dokarvam do prosheska toyaga* (lit. 'bring to a beggar's stick') 'beggar, pauperise'. Another frequent variant in patterns with definite nominal dependents features an expletive possessive V + expl:poss + obj – (ro) *își rupe spatele* 'break one's back' or reflexive possessive clitic V + det + obj – (bg) *iztarvavam si nervite* (lit. 'drop one's nerves') 'lose one's temper'. In both languages the possessive clitic occurs only with definite nouns or noun phrases.

Another relatively frequent pattern, as shown in Table 2, is the one containing an object that is restricted to the singular (definite and indefinite) forms: see the examples (ro) *avea încredere* 'have trust' – (bg) *imam vyara* 'have faith'. A plural object (e.g., (ro) *închide ochii*) is more rarely found in the Romanian data as compared with the singular, although in Bulgarian the patterns with plural definite complements are quite well represented: see examples (bg) *darvam kontsite* and (bg) *hodya po nervite*.

In Bulgarian, unrestricted objects/modifiers are also represented to a certain extent. Examples such as (bg) *iznasyam lektsia* and (bg) *vzemam prisartse* show patterns with a nominal complement unrestricted for number and definiteness, or an adverbial modifier, that is unrestricted for the category of degree (comparative, superlative), which is possible for some MWEs. In Romanian, such examples could not be found in the dataset.

Table 2: The most frequent morphosyntactic restrictions on dependents found with VMWEs in Bulgarian and/or Romanian (literal translation is provided only when it differs from the English equivalent).

Restrictions	Romanian example	Bulgarian example
Number = sg Def = indf V + obj	<i>lua parte</i> 'take part'	<i>vzemam uchastie</i> 'take part'
Number = sg Def = indf V + [case + obl]	<i>pune pe fugă</i> lit. 'put on running' 'oust, cause to flee'	<i>obrashtam v byagstvo</i> lit. 'turn into flight' 'oust, cause to flee'
Number = sg Def = def V + obj	<i>atrage atenția</i> lit. 'attract attention' 'call attention'	<i>nasochvam vnimanieto</i> lit. 'direct attention' 'call attention'
Number = sg Def = def V + [case + obl]	<i>sta la baza</i> lit. 'stand in the base' 'underlie'	<i>lezha v osnovata</i> lit. 'lie in the base' 'underlie'
Number = sg V + obj	<i>avea încredere</i> lit. 'have trust' 'trust'	<i>imam vyara</i> lit. 'have faith' 'trust'
Number = pl Def = def V + obj	<i>închide ochii</i> lit. 'close eyes' 'turn a blind eye'	<i>darbam kontsitate</i> 'pull strings'
Number = pl Def = def V + [case + obl]	<i>fi cu ochii</i> lit. 'be with eyes' 'keep an eye on'	<i>hodya po nervite</i> lit. 'walk on the nerves' 'madden'
Unrestricted V + obj	–	<i>iznasyam lektsia</i> lit. 'present a lecture' 'lecture'
Unrestricted V + advmod	–	<i>vzemam prisartse</i> 'take to heart'
Def = def V+ expl:poss/det+obj	<i>își rupe spatele</i> 'break one's back'	<i>prosyva si belyata</i> lit. 'beg for my own trouble' 'ask for trouble'

4.3.3 Valence frames

Another important aspect of the syntactic description of VMWEs is represented by their valence frames, which we encode by the use of the following conventions. First, they are formulated as UD relations: for each MWE, we define the types of relations it establishes within a sentence to ensure its grammatical correctness. For example, the MWE *kick the bucket* has a valence frame containing only the subject, i.e., *nsubj*.

The valence frames can contain obligatory, as well as optional relations. The difference between them is that the latter can be absent from the sentence without affecting its grammatical correctness: consider the sentence in (8):

- (8) *Regizorul i -a dus de nas pe spectatori cu un scurtmetraj.* (ro)
 Director them has taken of nose on audience with a short-film
 lit. ‘The director lead the audience by the nose with a short film.’
 ‘The director pulled the wool over the audience’s eyes with a short film.’

The subject *Regizorul* and the object *spectatori* are obligatory relations, but the prepositional object *cu un scurtmetraj* is optional. The optional nature of a relation is marked by means of round brackets around it; thus the valence frame for the VMWE in (8) is: *nsubj, obj, (case{cu}, obl)*.

Third, lexical restrictions on the form of prepositions or markers are rendered between curly brackets immediately after the relevant relation, case and mark, respectively: e.g., *case{cu}* in the frame presented for example (8).

Fourth, alternative valences are separated by a slash. For example, if two different prepositions occur after a VMWE, they are listed as values of the respective relation in the manner described: *case{împotriva/asupra}*.

Fifth, whenever an alternative consists of at least two elements (e.g., relations, forms, etc.), they are grouped together within square brackets: for example, (ro) *da drumul* (lit. ‘give way’) ‘let go’ can take either a prepositional object with the preposition *la* or an indirect object; this is represented as follows: *[case{la}, obl]/iobj*.

Table 3 shows the most frequent valence frames characterising VMWEs in the Bulgarian and Romanian datasets. Most of the encoded valences describe personal verb constructions, thus they require a subject in the frame, unless it is part of the expression, which happens rarely, as mentioned above.

When correlating the PARSEME VMWE types with their valence frames, we notice the following. Besides the subject, the valence frames of all expressions of the type *LVC.cause* have an obligatory object. This is in line with the definition of this type in PARSEME, according to which the noun in the *LVC.cause* “has

Table 3: The most frequent valence frames in the two languages.

Valence frame	Romanian example	Bulgarian example
nsubj	<i>o lua la goană</i> lit. ‘her take at rush’ ‘break away’	<i>hvashtam gorata</i> lit. ‘take the wood’ ‘take to the woods’
nsubj, obj	<i>aduce în sapă de lemn</i> lit. ‘bring in hoe of wood’ ‘pauperise’	<i>dokarvam do prosiya</i> lit. ‘bring to beggary’ ‘pauperise’
nsubj, iobj	<i>da frâu liber</i> lit. ‘give rein free’ ‘unleash’	<i>davam volya</i> lit. ‘give freedom’ ‘unleash’
nsubj, case, obl	<i>da piept</i> lit. ‘give breast’ ‘confront’	<i>varvya v krak</i> lit. ‘walk in step’ ‘keep pace’
nsubj, [case, obl] / [mark, ccomp]	<i>da seamă</i> lit. ‘give count’ ‘be responsible for’	<i>namiram sili</i> lit. ‘find strength’ ‘take heart’

semantic arguments expressed as non-subject elements in the sentence”.²¹ E.g., (ro) *pune în circulație* (lit. ‘put in circulation’) ‘issue’ has the internal structure V + [case + obl] and the valence frame nsubj, obj, where the obl has the obj as a semantic argument – see the example: *Banca pune banii în circulație* (lit. ‘Bank puts money in circulation’) ‘The bank issues money’, in which *money* is the semantic argument of *circulație*.

The valence frames of VMWEs of the types LVC.full and VID may contain only the subject or the subject and a nominal (obj, iobj or obl) or a clause: here are some examples: (a) VID (ro) *prinde inimă* (lit. ‘catch heart’) ‘cheer up’ takes only a subject: *Copilul a prins inimă* ‘The child cheered up’; (b) VID (ro) *purta sâmbetele* (lit. ‘bear Saturdays’) ‘bear ill will’ takes a subject and an indirect object: *Bărbatul îi purta sâmbetele soacrei sale* ‘The man was bearing his mother-in-law ill will’; (c) VID (ro) *cădea de acord* (lit. ‘fall of agreement’) ‘reach agreement’ takes a subject, an oblique indicating the person with whom agreement is achieved, and a subordinate clause or a prepositional phrase indicating the matter which was the subject of discussion: *Avocatul a căzut de acord cu*

²¹https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_Cross-lingual_tests/020_Light-verb_constructions__LB_LVC_RB_

clientul [asupra onorariului]/[cât să îl plătească] ‘The lawyer has reached agreement with his client [on the fee]/[how much to pay him’; (d) LVC.full (ro) *avea încredere* (lit. ‘have trust’) ‘trust’ takes a subject, an oblique denoting the person who the subject trusts, and a subordinate clause indicating with respect to what the subject trusts the other person: *Bărbatul are încredere în avocat că va câștiga procesul* ‘The man trusts the lawyer that he will win the trial’.

In addition to these, both languages display valence patterns where one of the elements is an obligatory nmod or complement that usually enters the relation obj or obl with the verb, e.g., (ro) *sta la baza* (lit. ‘stand at the base’) and (bg) *lezha v osnovata* (lit. ‘lie in the base’) where the obliques (ro) *baza* and (bg) *osnovata* need a nominal modifier to form a grammatical sentence. These may also be possessive phrases, e.g., (bg) *hodya po nervite* + nmod: *na nyakogo* (lit. ‘walk on the nerves + nmod: of someone’) ‘madden’.

Empty valence frames are also possible where the VMWEs are headed by impersonal verbs and they do not have obligatory complements or modifiers. In Romanian, this is the case of weather expressions, such as (ro) *plouă cu găleata* (lit. ‘rains with bucket’) ‘it is raining cats and dogs’. The corresponding Bulgarian expression (bg) *vali kato iz vedro*, with the same meaning, may be headed by an impersonal or by a personal verb and thus takes alternatively either an empty or an nsubj frame.

4.3.4 Word order variation

Both languages are characterised by a relatively free word order. The manual analysis of the VMWEs and the validation of this linguistic introspection using large corpora show that most VMWEs are no exception to this general rule. Here is an example of a LVC.full in Romanian (9) and in Bulgarian (10) showing this free word order:

- (9) a. **Luăm parte** la concert. (ro)
 Take part at concert
 ‘We take part in the concert.’
- b. **Parte luăm** la concert. (ro)
 Part take at concert
 ‘We take part in the concert.’
- (10) a. В концерта **взеха участие** известни изпълнители. (bg)
 V kontserta **vzеха uchastie** izvestni izpalniteli.
 In concert-DEF took part famous performers.
 ‘Famous performers took part in the concert.’

- b. В концерта участие взеха известни изпълнители. (bg)
V kontserta uchastie vzeha izvestni izpalniteli.
In concert-DEF part took famous performers.
'Famous performers took part in the concert.'

However, when (some) constraints exist with respect to the word order of components or only of some of them, they are clearly marked in the entry of the respective VMWE. Such examples include: (ro) *arunca praf în ochi* (lit. 'throw dust in eyes') 'pull the wool over one's eyes', in which the noun phrase (*praf*) and the prepositional phrase (*în ochi*) always occur in this order, and the verb can be moved after them, thus resulting in an emphatic construction. A relevant example is (bg) *mnogo mi znae ustata* (much my knows mouth-DEF, lit. 'my mouth knows a lot') 'have a big mouth'. The normal word order of the MWE is an emphatic one with the *advmod* first and the *nsubj* last instead of the neutral sentential order *nsubj* + *det* + *V* + *advmod*. Although even in this case different word order variants are possible, some of them such as the ones where the *advmod* follows the *V* or the *V* follows the *nsubj* are very rare and we mark them as such.

4.3.5 Intervening elements

Another syntactic characteristic of VMWEs in the two languages is the possibility for (sequences of) words that do not belong to the expression to occur between its components. This is a consequence of the relatively free word order characterising Bulgarian and Romanian. Such an example is: (ro) *Învăţ adesea, cu drag, o poezie pe de rost* (lit. 'Learn often, with pleasure, a poem by heart'). A few words occur within the VID *învăţa pe de rost* 'learn by heart': a frequency adverb (*adesea* 'often'), a manner prepositional phrase (*cu drag* 'with pleasure') and the direct object (*o poezie* 'a poem'). The first two are not part of the valence frame, whereas the last one is. We take the stance that by default the VMWEs obey the general rules of the language in question so that peculiarities resulting from the free word order need not be marked in any way.

However, there are also cases in which the possibility for intervening elements is blocked. Such an example is: compare (ro) *Stă cu mâinile adesea în sân* 'She often stays with her arms crossed' with *Stă adesea cu mâinile în sân* 'She often stays doing nothing'. The former example shows that it is not possible to insert the frequency adverb *adesea* 'often' between the two prepositional phrases of the VID and keep the non-compositional meaning (hence, the status of VID),

whereas the latter shows that this insertion is possible between the verb and the first prepositional phrase.

For Bulgarian, we note that in some cases external elements may be blocked between a dependent's modifier and its head, when both are part of the idiom (V + [case + amod + obj]), e.g. (bg) *stoya sas skrasteni ratse* (lit. 'sit with crossed arms') 'sit back, sit by'. In this case, the occurrence of such element signals that the phrase has a literal reading, as in (bg) *Toy stoeshе sas skrasteni otpred ratse* 'He stood with his arms crossed in front of his body'.

There are also cases where parts of the VMWE are themselves idiomatic and thus do not allow intervening elements. Consider the example (bg) *varvya v krak* 'keep in step' whose dependent *v krak* 'in step' functions as an idiomatic expression outside the VID, and therefore the noun cannot be modified.

Wherever we establish restrictions on the occurrence of intervening elements between the components of a VMWE, the lexicon entry clearly states the components between which such an insertion is blocked.

Theoretically, the intervening elements may belong to a particular part of speech, may be forms of a particular lexeme or lexemes, etc. At the current stage of the development of the MWE lexicon, we prefer to collect evidence of various types of idiosyncrasies whose tackling may be dealt with at present, or may be deferred to a later moment. One of the focuses of this part of our work are the cases that diverge from the regular syntactic and linearisation rules of the language under study. Currently, this description involves the specification of POS tags that are disallowed. In the above case, the VID (bg) *varvya v krak* 'keep in step' does not allow the modification of the noun, although the rules of Bulgarian license the adjectival modification of nominals in prepositional phrases.

4.4 Semantic description

The lexicon design proposed in this chapter falls in line with the trend of describing MWEs in dedicated lexicons that provide various types of linguistic information referring to the MWE and its components and may be employed in MWE recognition related tasks. Due to the fact that BulNet and RoWN are aligned to each other, to WordNet and to any other wordnet mapped to it (see §3.1), we make use of the rich semantic description provided in the WordNet, added from additional resources or supplied manually by the teams developing BulNet and RoWN. The use of WordNet further supports the multilingual dimension of the described resource through the possibility of directly deriving the relevant semantic description available for other languages.

The main components of the semantic description incorporated herein are: a definition (called gloss), a set of semantic relations to other WordNet concepts, usage examples, stylistic and connotation information.

4.4.1 Lexicographic definition

The lexicographic description of MWEs in the form of definitions was employed by various authors of MWE resources, including close non-MWE paraphrases (cf. §2). The use of definitions aims not only at documenting the meaning of a MWE, but also at distinguishing the particular sense from other senses of the same MWE lemma, thus accounting for polysemy.

The lexicographic definition adopted in WordNet and in the lexicon describes concepts regardless of the structure of the units that lexicalise them (single words or MWEs). Thus each MWE shares a definition with the remaining synonyms in the relevant synset in both languages, with the WordNet gloss serving as an intermediary.

4.4.2 Stylistic and/or register information

The inventories for encoding stylistic/register information in MWE resources are usually subsets of those adopted in standard dictionaries (§2.7). Note that while stylistic remarks are usually assigned to an entry, which means that they characterise all the occurrences of the respective lexical unit, Fellbaum & Geyken (2005) assign the labels to usages, thus accounting for the fact that the same idiom may have different stylistic features depending on the context.

In the model adopted, we assign stylistic/register information as a permanent value attached to a MWE, using one or more labels, established in the lexicographic practice and adopted in the BulNet for both single and MWE lexemes: “colloquial”, “slang”, “literary”, “figurative”, “dialect”, “obsolete”, “pejorative”. The values were assigned to the RoWN counterparts and reviewed manually, as lexical items describing the same concept may differ stylistically. Thus, the corresponding VIDs (bg) *davam pet pari* (lit. ‘give five paras’) and *davam puknata para* (lit. ‘give broken para’) and (ro) *da doi bani* (lit. ‘give two coins’) ‘give a hang’ are marked as “colloquial”, whereas (ro) *da două parale* (lit. ‘give two paras’) having the same meaning but pertaining to a different register is marked as “literary”.

4.4.3 Connotation

We include connotative information which is automatically assigned to BulNet and RoWN from SentiWordNet (Baccianella et al. 2010). This is an open lexical

resource designed for supporting sentiment classification and opinion mining applications which resulted from the automatic annotation of all the synsets in WordNet with one of three possible values: positive (between 0.00 and 1.00), negative (between -1.00 and 0.00) and neutral (0.00). The sentiment values were assigned to BulNet and RoWN as part of previously implemented tasks.

In our current work, we undertook a check of the values at the level of individual VMWEs (not the level of the synset), as different literals may have different connotation. For instance, the colloquial (bg) *hvarlyam prah v ochite* and (ro) *arunca praf in ochi* ('throw dust in the eyes') have negative connotation, but the synset was assigned a positive value of 0.5. We marked where the connotation value assigned from WordNet were reconsidered in our resource.

4.4.4 Semantic relations

Another trend in MWE lexicon crafting was to integrate MWEs into the lexical system of the language as individual entities, while accounting for their morphological, syntactic and semantic properties. This integration may involve the encoding of various relations to other single and MWE lexemes (§2).

By virtue of their integration in the WordNet's structure, the VMWE in the devised lexicon are explicitly associated to their synonyms (i.e., the remaining synset members, both single words and MWEs), see Figure 2. Through their membership in synsets, VMWEs are also connected to other synsets in WordNet via a number of conceptual-semantic relations – hypernymy (and its inverse hyponymy), holonymy (and its inverse meronymy), etc. – and/or lexical relations, e.g., antonymy (Miller 1995, Fellbaum 1998b).

The Bulgarian and Romanian MWEs in the target synset are connected to their hypernym (also containing MWE literals in Bulgarian). In addition, WordNet includes derivational relations (marked as *eng_derivative*), part of which are assigned semantic values that denote various roles in the situation described, eventualities or properties, i.e., the so-called morphosemantic links (Fellbaum et al. 2009). Derivational relations require validation as they might not be true across languages, e.g., (bg) *magazin:1* and (ro) *magazin:1; prāvālie:1* ('shop') are not derivationally related to the target synset. Their semantic values, however, are considered to be language-independent. In Figure 2, such relations are: *has_location* that connects the target synset to the location where it takes place; *has_agent* – pointing to the invariant agent (a person who shops); *has_event* – the act of doing shopping. Another relation, *category_domain*, describes the domain to which a synset pertains (if relevant). In this case it relates the target synset to the domain of commerce.

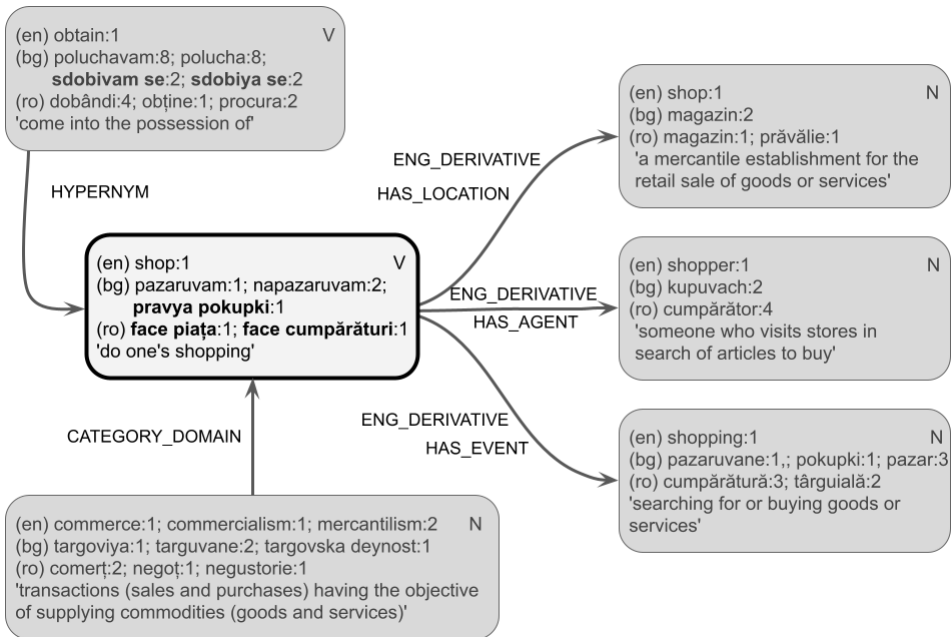


Figure 2: Synset relations within WordNet.

4.5 Derivational information

MWEs can be bases for derivation in both Romanian and Bulgarian, but this property was not consistently accounted for in WordNet.²² Barbu Mititelu & Leseva (2018) showed that derivation of MWEs can result into both other MWEs and one-word compounds; the authors also analysed some syntactic patterns identified in the derivation of VMWEs extracted from two lexicons of MWEs in Bulgarian and Romanian. However, the VMWEs in our lexicon display only the derivational relations between two MWEs.

We also investigated the derivational potential of the VMWEs included in the lexicon. Our datasets do not coincide with those used by Barbu Mititelu & Leseva (2018), although a certain overlap is naturally possible. However, after the manual investigation of the derivational possibilities of the VMWEs in BulNet

²²Note that we cannot claim that the discussed patterns are indeed resulting from a process of derivation that occurred in the language history. Rather, we mean that there are multiword formations that are semantically and structurally related to VMWEs and that those formations involve the employment of some mechanism of derivation (or even inflection) concerning one or more of the elements of the respective VMWE, as well as internal syntactic restructuring.

and RoWN, we could confirm the patterns enumerated there.²³ Table 4 shows the syntactic patterns involved in the VMWEs derivation, alongside examples for each language in which they are found. Derivational patterns with the same syntactic transformation, but involving different semantics, are presented as distinct patterns (e.g., $V + \text{obj} > N_V\text{-derived} + \text{case} + \text{nmod}$ for deriving Event or Agent). The head of the derivation is marked by boldface.

The data shows a vast number of nouns designating events, which is in line with the findings by Barbu Mititelu & Leseva (2018), while derivation involving a result pertaining to other semantic types is less numerous. The semantic labels provided in Table 4 are mostly based on the inventory analysed by Barbu Mititelu & Leseva (2018).

In light of the most represented syntactic patterns in the datasets, the primary bases for forming the most productive type – event deverbal – are VMWEs exhibiting the relations $V + \text{obj}$ and $V + [\text{case} + \text{obl}]$. In addition to expressing the VMWE complement as a prepositional modifier in the resulting nominalisation, Romanian exhibits a pattern where the VMWE complement is turned into a genitive modifier, which the Bulgarian language does not allow for.

From the same syntactic patterns, but involving a different (e.g., agentive) suffix in the derivation of the deverbal noun, we obtain Agents (bg) *perach na pari* ‘brainwasher’, Patients (ro) *muritor de foame* ‘very poor person’, etc.

VMWEs exhibiting the $V + \text{obj}$ relation allow the formation of noun expressions (NMWEs) whose head is the object of the VMWE modified by a participial (adjective) – a past (passive) participle, cf. the examples in Table 4: (bg) *promiya mozaka > promit mozak* and (ro) *trage sfori > sfori trase*. The meaning is resultative and aligns with such examples in English: (en) *close the door > closed door*, *break the heart > broken heart*.

VMWEs exhibiting both the relations $V + \text{obj}$ and $V + \text{case} + \text{obl}$ regularly correspond to formations headed by a participle of the head verb in the VMWE. In Bulgarian different participles take part in this process: present (active) participles, e.g. (bg) *smrazyavam krvta > smrazyavasht krvta* ‘curdle the blood’ > ‘curdling the blood’, ‘blood-curdling’; past active participles: (bg) *umiram ot glad > umryal ot glad* ‘die of hunger’ > ‘dead of hunger’, ‘starved’; past passive participles: (bg) *vlyubya se do ushi > vlyuben do ushi* ‘fall in love to the ears’ > ‘fallen in love to the ears’, ‘be head over heels in love’ > ‘head over heels in love’. Some of them become established in the language and are converted to adjectives,

²³The patterns presented by Barbu Mititelu & Leseva (2018) are described in terms of dependency grammar, but using syntactic functions such as subject, complements, adjuncts. The confirmation of those patterns was possible by converting them into the UD format.

Table 4: The most frequent syntactic patterns involved in the VMWE-to-OtherPOS-MWE derivation.

Romanian examples	Bulgarian examples	
V + case + obl > N_V-derived + case + nmod		Event
<i>ieși la iveală</i> > <i>ieșirea la iveală</i> 'exit at apparition' > 'exit (N) at apparition' 'come to light' > 'coming to light'	<i>umiram ot glad</i> > <i>umirane ot glad</i> 'die of hunger' > 'an act of dying of hunger' 'starve' > 'starving, starvation'	
V + obj > N_V-derived + case + nmod		Event
<i>spăla bani</i> > <i>spălare de bani</i> 'laundry money' > 'laundrying of money', 'money laundrying'	<i>pera pari</i> > <i>prane na pari</i> 'laundry money' > 'laundrying of money', 'money laundrying'	
V + obj > N_V-derived + nmod		Event
<i>spăla creierul</i> > <i>spălarea creierului</i> 'brainwash' > 'brainwashing'	-	
V + obj > N_V-derived + case + nmod		Agent
<i>spăla creierul</i> > <i>spălător de creiere</i> 'brainwash' > 'brainwasher'	<i>promivam mozaka</i> > <i>promivach na mozatsi</i> 'brainwash' > 'brainwasher'	
V + obj > ADJ_V-derived + N_{obj}		Result
<i>trage sfori</i> > <i>sfori trase</i> 'pull strings' > 'pulled strings'	<i>promiya mozaka</i> > <i>promit mozak</i> 'brainwash' > 'a brainwashed brain'	
V + case + obl > ADJ_V-derived + case + obl		Characteristic
<i>muri de foame</i> > <i>mort de foame</i> 'die of hunger' > 'dead of hunger' 'starve' > 'starving'	<i>umra ot glad</i> > <i>umryal ot glad</i> 'die of hunger' > 'dead of hunger' 'starve' > 'starving'	
<i>spăla creierul</i> > <i>spălat pe creier</i> 'brainwash' > 'brainwashed'		
V + case + obl > ADJ_V-derived + case + obl		Characteristic
<i>scoate din minți</i> > <i>scoatere din minți</i> 'take-out from minds' > 'taking-out from minds' 'madden' > 'maddenning'	<i>umiram ot glad</i> > <i>umirasht ot glad</i> 'die of hunger' > 'dying of hunger' 'starve' > 'starving'	

whereas others are used in context but are not established as lexicographic units. Nevertheless, such constructions need to be described both from the perspective of generation, as they are formed on the basis of VMWEs having a certain syntactic structure and morphological properties according to certain rules, and recognition (being able to associate a relevant string of forms as related to the source VMWE).

With respect to derivation, the Romanian dataset contains a large number of VMWEs which are bases for derived nominal MWEs by means of conversion applied to the supine verb of the base VMWE. For example, (ro) *da socoteală* lit. ‘give payoff’ ‘answer for’ is the base for *datul socotelii*: the derived nominal MWE is obtained from the base MWE by converting the supine of the verb *da*, namely *dat*, into a noun, shown here by adding the definite article *-(u)l* to it. Equally often we find cases when the participle of the verb allows for the derivation of an adjectival MWE from the verbal one, also by means of conversion: e.g., (ro) *trage pe sfoară* lit. ‘pull on rope’ ‘play a trick on’ is the base for *tras pe sfoară*: the derived adjectival MWE is obtained from the base MWE by conversion of the supine of the verb *trage*, namely *tras*, into an adjective, which is a frequent phenomenon in Romanian.

4.6 Visualisation and basic query interface

Figure 3 shows the basic visual interface that allows access to and queries on the dataset. There are several filtering parameters: (i) the type of the VMWE (All, VID, LVC); (ii) word order variability; (iii) syntactic flexibility – whether the VMWE allows its components to be modified; (iv) stylistic register of the MWE; (v) structure of the VMWE – syntactic patterns according to the UD scheme; (vi) search terms in either Bulgarian and/or Romanian VMWEs or abstract lemmas. The result of the filtering is a list of all VMWE pairs that match the filtering criteria. Each VMWE pair is first identified by its synset ID and WordNet definition. If more than one VMWE pairs are available for a given synset, the user can select among possible Bulgarian-Romanian literal pairs to align and compare. Upon selection, the pair of VMWEs is presented in parallel for Bulgarian and Romanian (see Figure 4) with the features outlined in §4.1–§4.5.

Search instructions

Type: All VID LVC

Word order: All Frozen Limited Free

Allowing modifiers: All No modifiers Accepting Modifiers

Register: All Marked as colloquial Other marked

Syntactic pattern:

- All
- V + [amod + case + obl]
- V + [amod + obj]
- V + [case + advmod]
- V + [case + amod + obl]
- V + [case + case + obl]
- V + [case + obj]
- V + [nummod + obj]
- V + [obj (pronoun) + amod]
- V + advmod
- V + expl:poss + [case + obl]
- V + expl:poss + obj
- V + expl:poss + obj + xcomp
- V + expl:pv + [case + obl]
- V + expl:pv + [obl + case + obl]
- V + obj
- V + obj + [case + obl]
- V + obl (short dative pron) + [case + obl]
- cop + [case + ROOT]
- cop + [case + nummod + ROOT]
- cop + [case+ ROOT]
- cop + advmod
- neg + V + xcomp

Search by component in the MWE or abstract lemma of the MWE:

Word (bg) Word (ro)

Figure 3: Search interface to filter MWE data.

- eng-30-00839194-v
conceal one's true motives from especially by elaborately feigning good intentions so as to gain an end
- bg**
ro
- хвърля прах в очите
 - хвърлям прах в очите
 - arunca praf în ochi
 - duce de nas
 - trage pe sfoară

Feature	BG	RO
WORDNET ID	eng-30-00839194-v	eng-30-00839194-v
Literal	хвърлям прах в очите	arunca praf în ochi
PARSEME TYPE	VID	VID
Definition	карам някого да мисли или прави нещо в моя угода, прикривайки реалните факти, мотиви или цели	A induce pe cineva în eroare, printr-o viclenie sau printr-o minciună, pentru a trage un folos sau pentru a se amuza
MWE Lemma	хвърлям прах в очите	arunca praf în ochi
MWE Abstract Lemma	хвърлям прах в око	arunca praf în ochi
Aspect (BG only)	IMPERF	-
Regular morphosyntactic representation		NO
Restrictions on the verbal head	NO	NO
Restrictions on dependents	2_fixed: N = s; D = 0 & 3_fixed & 4_fixed: N = p; D = d	praf - only singular, only without article; ochi - only without article
Internal structure (in UD format)	V + obj + [case + obl]	V + obj + [case + obl]
Valences	nsubj, nmod:poss	nsubj, iobj
Word order restrictions		praf în ochi - as is
Intervening words blocked		praf în ochi
Register	colloquial	colloquial
Sentiment - pos	0	0
Sentiment - neg	0.5	0.5
Derivation	хвърляне на прах в очите	aruncatul prafului în ochi; aruncarea prafului în ochi

Figure 4: Visualisation of aligned bilingual VMWEs.

5 Discussion, conclusions, and future work

We consider the important aspects of our work to be (i) its focus on languages other than English, and (ii) the use of a common framework for an in-depth linguistic description of VMWEs. Bulgarian and Romanian are morphologically richer languages than English and belong to different families (Slavic and Romance, respectively). The description of VMWEs in these two languages is made in a multilingual landscape offered by aligned wordnets. Using of a common framework for an in-depth linguistic description of VMWEs allows for highlighting both similarities and differences between the MWEs in the two languages. Moreover, this framework is encoded in a transparent, flexible, expressively capable, versatile and friendly way (Lichte et al. 2019).

Our lexicon is rooted in WordNet: the organisation principles therein explain the work methodology and the representation of information. Thus, for a MWE, we do not encode a list of lexemes that can substitute components in an expression, as is the case with some other such lexicons (see §2). Whenever such substitutions are possible, the whole expression is encoded as a different literal occurring in the same synset as its synonyms (thus, labelled with different literal IDs, see §4.1). One such example is the pair (ro) *da doi bani – da două parale* ‘give a hang’, which differ in their last component: *ban* is a current unit of money, while *para* is an older one, not used anymore. An argument in favor of the distinct treatment of lexical variants is that, other differences aside, as we showed earlier, the two MWEs belong to different lexical registers – one is colloquial and the other is literary.

There are also cases when two expressions vary by means of one component that is added to offer emphasis to the expression in use: see the pair (ro) *își da silința – își da toată silința* ‘do one’s best’, which differ only in the determiner *toată* ‘all’ added to the direct object of the verb, thus making it more emphatic. This affects the communicative status of the different variants and may determine the choice of one over the other in a context, the preference of different equivalents or translations in other languages, etc.

Another consequence of including in the lexicon MWEs from wordnets is that no relationship is encoded between an expression and the entries for its components, i.e., the synset(s) to which the MWE belongs and the synsets to which its components belong (unlike traditional thesauri where MWEs often appear under one or more of its components). Each word sense and each MWE sense are separately encoded. However, by means of the relations in the networks, when any semantic relation exists between one meaning of a component and the meaning of a MWE of which it is a component, then this (close or distant) relation can

be retrieved by traversing the edges starting from one synset and reaching the other one.

The multilingual dimension of the resource presented here springs from the fact that the Bulgarian-Romanian lexicon exploits the alignment between the two wordnets, thus being a resource on top of two linked monolingual ones. The alignment was possible via Princeton WordNet and this actually opens the way to alignment to any other such lexicon either built on top of other wordnets or linked to them. A possible future development towards the multilingual extension would be to employ a large-scale densely populated resource providing access to aligned MWE entities such as BabelNet (Navigli & Ponzetto 2012).

Lichte et al. (2019) discuss what they call general virtues of MWE encoding, namely *transparency*, *flexibility*, *power to generalise*, *implementation friendliness*, *electronic versatility*, as prerequisites for a lexical resource. *Transparency* concerns the ability of the human user to map the encoding back to the source set of lexical properties, i.e. the simplicity of the encoding of linguistic features and the straightforwardness of their interpretation by novices or non-expert users. *Flexibility* is the adaptability of a format to dealing with unforeseen properties or changes in properties. The *power to generalise* allows the user to group properties and assign them collectively, thus avoiding redundancies and errors. The *implementation friendliness* relates to the existence of tools that assist a human user with encoding or its validation. *Electronic versatility* describes the ease of converting the lexical encoding into a lexical resource, in particular, the existence of conversion tools or the possibility to produce them.

To ensure the *transparency* of the encoding, we adopted a straightforward link between the linguistic properties and their values. The field names serving to encode the properties are both easy to encode manually and to interpret. The basic tabular format of the template used to describe each MWE component facilitates the adaptation to new or unforeseen properties, thus ensuring the *flexibility* of the data encoding. New (categories of) fields and values may be defined as appropriate when needed and added to the predefined VMWE description template. This is especially relevant with respect to language-specific features (e.g., verb aspect in Bulgarian), as it allows the two teams to work independently. The unified description of the data for each language enabled us to consider two aspects of the *power to generalise*: (i) the possibility to identify and extract linguistic regularities, including groups of relevant properties in the VMWEs that share them, thus identifying possible classes of VMWEs with similar characteristics (from a certain perspective); and (ii) the possibility to look into linguistic regularities or shared features between the languages as well as to extract semantic, structural, etc. correspondences between VMWEs in them. *Implementation friendliness* and *electronic versatility* stem from the simple form in which the data are described.

Currently, we did not use a particular tool, but the explicitness of the format and the encoding of features makes it easy to convert to various formats according to the relevant requirements of the existing tools.

Adopting the same work methodology made it possible for the teams to work independently from each other using a predefined template that includes the relevant linguistic features (on the basis of previous data analysis) and expanding it to new features when the need arises. The model is thus adaptable to languages that share similar linguistic properties, possibly to genetically and/or typologically related ones.

Future work will aim at the enrichment of the monolingual lexicons with descriptions of the VMWEs that are in the individual wordnets, as for now we created entries only for those that are mutually equivalent VMWEs in the two languages. The further development of the two wordnets will allow for the identification of other (V)MWEs equivalents, thus enriching the bilingual lexicon and extending it to MWEs of other parts of speech.

Syntactic transformations have not been tackled yet in our resource. As most of them show the regular syntactic behavior of free phrases, we have decided, during the next stage of our work, to start marking the cases where a certain transformation is impossible and proceed to describe the conditions for blocking. This will be implemented in a manner similar to the encoding of morphosyntactic restrictions, i.e. by defining a relevant field ‘Syntactic transformations’ and listing the restrictions using a predefined list of the names of the transformations as values. A further, more in-depth treatment of syntactic transformations will depend on the analysis of the data after we have collected them.

As corpora annotated with VMWEs exist for both Romanian (Barbu Mititelu et al. 2019a) and Bulgarian (Koeva et al. 2012), associating the lexicon entries with relevant corpora occurrences is a natural next step that would contribute a syntagmatic dimension to the resource.

Abbreviations

BulNet	Bulgarian WordNet	N	noun
HPSG	Head-driven Phrase Structure Grammar	NLP	Natural Language Processing
IAV	inherently adpositional verb	NMWE	noun multiword expressions
IRV	inherently reflexive verb	RoWN	Romanian WordNet
LFG	Lexical-Functional Grammar	UD	Universal Dependencies
LVC	light verb constructions	V	verb
MWE	multiword expression	VID	verbal idiom
		VMWE	verbal multiword expression

Acknowledgements

The authors are grateful to the anonymous reviewers and to the editors of this volume for their remarks on the previous versions of this chapter, which helped improving it.

References

- Autelli, Erica. 2020. Phrasemes in Genoese and Genoese-Italian lexicography. In Joanna Szerszunowicz & Eva Gorlewska (eds.), *Applied linguistics perspectives on reproducible multiword units: Foreign language teaching and lexicography* (Intercontinental Dialogue on Phraseology 8), 101–127. Białystok: University of Białystok Publishing House. DOI: 10.1007/978-3-642-30910-6_12.
- Baccianella, Stefano, Andrea Esuli & Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh international Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.
- Barbu Mititelu, Verginica, Mihaela Cristescu & Mihaela Onofrei. 2019a. The Romanian corpus annotated with verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 13–21. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5103.
- Barbu Mititelu, Verginica & Svetlozara Leseva. 2018. Derivation in the domain of multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 215–246. Language Science Press. DOI: 10.5281/zenodo.1182601.
- Barbu Mititelu, Verginica, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova & Maria Todorova. 2019b. Hear about verbal multiword expressions in the Bulgarian and the Romanian Wordnets straight from the horse's mouth. In *Proceedings of the joint workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 2–12. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5102.
- Bresnan, Joan. 1978. A realistic transformational grammar. In Morris Halle, Joan Bresnan & George A. Miller (eds.), *Linguistic Theory and Psychological Reality*, 1–59. MIT Press.
- Chiarcos, Christian, Maxim Ionov, Elena-Simona Apostol, Katerina Gkirtzou, Besim Kabashi, Anas Fahad Khan & Ciprian-Octavian Truică. 2024. Multiword expressions, collocations and the OntoLex vocabulary. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources:*

- Linguistic, lexicographic, and computational perspectives*, 187–227. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998641.
- Dalrymple, Mary (ed.). 2023. *Handbook of Lexical Functional Grammar* (Empirically Oriented Theoretical Morphology and Syntax 13). Berlin: Language Science Press. DOI: 10.5281/zenodo.10037797.
- de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.
- Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions*, 69–108. Language Science Press. DOI: 10.5281/zenodo.2579037.
- Fellbaum, Christiane. 1998a. Towards a representation of idioms in WordNet. In *Usage of WordNet in natural language processing systems*, 52–57. <https://aclanthology.org/W98-0707>.
- Fellbaum, Christiane (ed.). 1998b. *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.
- Fellbaum, Christiane & Alexander Geyken. 2005. Transforming a corpus into a lexical resource: The Berlin Idiom Project. *Revue française de linguistique appliquée* 10(2). 49–62. DOI: 10.3917/rfla.102.62.
- Fellbaum, Christiane, Anne Osherson & Peter E. Clark. 2009. Putting semantics into WordNet’s “morphosemantic” links. In Zygmunt Vetulani & Hans Uszkoreit (eds.), *Lecture notes in computer science*, 350–358. Berlin, Heidelberg: Springer.
- Fotopoulou, Aggeliki, Stella Markantonatou & Voula Giouli. 2014. Encoding MWEs in a conceptual lexicon. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 43–47. Gothenburg, Sweden: Association for Computational Linguistics.
- Fried, Mirjam & Jan-Ola Östman. 2004. Construction Grammar: A thumbnail sketch. In Mirjam Fried & Jan-Ola Östman (eds.), *Construction Grammar in a cross-language perspective*, 11–86. Amsterdam: John Benjamins.
- Giouli, Voula, Vera Pilitsidou & Hephestion Christopoulos. 2024. A FrameNet approach to deep semantics for MWEs. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 147–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998639.
- Grégoire, Nicole. 2007. Design and implementation of a lexicon of Dutch multiword expressions. In Nicole Gregoire, Stefan Evert & Su Nam Kim (eds.), *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 17–

24. Prague: Association for Computational Linguistics. <https://aclanthology.org/W07-1103>.
- Gross, Gaston. 1996. *Les expressions figées en français: Noms composés et autres locutions*. Paris: Ophrys.
- Gross, Maurice. 1975. *Méthodes en syntaxe: Régime des constructions complétives*. Paris: Hermann.
- Gross, Maurice. 1982. Une classification des phrases «figées» du français. *Revue québécoise de linguistique* 11(2). 151. DOI: 10.7202/602492ar.
- Al-Haj, Hassan, Alon Itai & Shuly Wintner. 2013. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography* 27. 130–170. DOI: 10.1093/ijl/ect036.
- Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová & Pavel Vondříčka. 2019. Lexical database of multiword expressions in Czech. In *Trudy mezhdunarodnoj konferencii Korpusnaja Lingvistika*, 9–16. Saint Petersburg, Russian Federation: Saint Petersburg University Press.
- Iñurrieta, Uxo, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka & Kepa Sarasola. 2018. Konbitzul: An MWE-specific database for Spanish-Basque. In *Proceedings of the eleventh international Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). <https://aclanthology.org/L18-1397>.
- Koeva, Svetla. 2010. Bulgarian WordNet: Current state, applications and prospects. In *Bulgarian-American dialogues*, 120–132. Sofia: Prof. M. Drinov Academic Publishing House.
- Koeva, Svetla. 2021. The Bulgarian WordNet: Structure and specific features. *Papers of the Bulgarian Academy of Sciences* 8(1). 47–70.
- Koeva, Svetla, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova & Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling* (1). 65–110. DOI: 10.15398/jlm.v0i1.33.
- Koeva, Svetla, Ivelina Stoyanova, Maria Todorova & Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of Bulgarian multiword expressions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the GLOBALEX 2016 workshop: Lexicographic Resources for Human Language Technology, LREC*, 86–95. Paris: European Language Resources Association (ELRA).

- Leseva, Svetlozara, Verginica Barbu Mititelu & Ivelina Stoyanova. 2020. It takes two to tango: Towards a multilingual MWE resource. In *Proceedings of the 4th international conference on Computational Linguistics in Bulgaria (CLIB 2020)*, 101–111. Sofia, Bulgaria: Department of Computational Linguistics, IBL – BAS. <https://aclanthology.org/2020.clib-1.11>.
- Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.
- Markantonatou, Stella, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chiril, Dimitrios Karamatskos, Nicolaos Valeontis & George Pavlidis. 2024. Description of Pomak within IDION: Challenges in the representation of verb multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 39–72. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998633.
- Markantonatou, Stella, Panagiotis Minos, George Zakis, Erasmia Koletti, Elpiniki Margariti & Emilia Stripeli. 2020. Idion (ιδίον): A lexicographic environment for the documentation of Greek idioms. In Stella Markantonatou & Anastasia Christofidou (eds.), *Multiword expressions in Greek: Deltio epistimonikis orologias ke neologismou*.
- Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019) at ACL 2019*, 130–134. Florence. DOI: 10.18653/v1/W19-5115.
- Mel’čuk, Igor. 1981. Meaning-text models: A recent trend in Soviet linguistics. *Annual Review of Anthropology* 10. 27–62.
- Mel’čuk, Igor. 2006. Explanatory combinatorial dictionary. In Giandomenico Sica (ed.), *Open problems in linguistics and lexicography*, 225–355. Monza, Italy: Polimetrica.
- Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.
- Monti, Johanna. 2014. An English-Italian MWE dictionary. In *Proceedings of the first Italian conference on Computational Linguistics CLiC-it 2014 and of the fourth international workshop EVALITA*, 265–269. Pisa: Pisa University Press.

- Müller, Stefan, Anne Abeillé, Robert D. Borsley & Jean-Pierre Koenig (eds.). 2021. *Head-Driven Phrase Structure Grammar: The handbook* (Empirically Oriented Theoretical Morphology and Syntax 9). Berlin: Language Science Press. DOI: 10.5281/zenodo.5543318.
- Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250. DOI: 10.1016/j.artint.2012.07.001.
- Odiijk, Jan. 2013. Identification and lexical representation of multiword expressions. In Peter Spyns & Jan Odiijk (eds.), *Essential speech and language technology for Dutch: results by the STEVIN programme*, 201–217. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-30910-6_12.
- Odiijk, Jan, Martin Kroon, Sheean Spoel, Ben Bonfil & Tijmen Baarda. 2024. MWE-Finder: Querying for multiword expressions in large Dutch text corpora. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 229–267. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998643.
- Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.
- Pollard, Carl & Ivan A. Sag. 1987. *Information-based syntax and semantics*, vol. 1: Fundamentals. Stanford: CSLI Publications.
- Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.
- Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk & Marcin Woliński. 2014a. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of workshop on lexical and grammatical resources for language processing*, 83–91. Dublin, Ireland: Association for Computational Linguistics & Dublin City University. DOI: 10.3115/v1/W14-5811.
- Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdziński. 2014b. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odiijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 2785–2792. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/279_Paper.pdf.

- Savary, Agata. 2008. Computational inflection of multi-word units: A contrastive study of lexical approaches. *Linguistic Issues in Language Technology* 1. DOI: 10.33011/lilt.v1i.1195.
- Savary, Agata, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archana Bhatia, Marie Candito, Polona Gantar, Uxoá Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze & Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 24–35. Dubrovnik, Croatia: Association for Computational Linguistics. <https://aclanthology.org/2023.mwe-1.6>.
- Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Van Der Plas, Behrang Qasemizadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471590.
- Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the joint workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91. Florence. DOI: 10.18653/v1/W19-5110.
- Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova & Federico Sangati. 2015. PARSEME: PARSing and Multiword Expressions within a European multilingual network. In *7th Language and Technology Conference: Human language technologies as a challenge for computer science and linguistics (LTC 2015)*. Poznań, Poland. <https://hal.archives-ouvertes.fr/hal-01223349>.
- Schafroth, Elmar. 2015. Italian phrasemes as constructions: How to understand and use them. *Journal of Social Sciences* 3(11). 317–337. DOI: 10.3844/jssp.2015.317.337.

- Shudo, Kosho, Akira Kurahone & Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 161–170. Portland, OR: Association for Computational Linguistics.
- Skoumalová, Hana, Marie Koprivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondříčka & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631.
- Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In Daniel Zeman & Jan Hajič (eds.), *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*, 197–207. Brussels, Belgium: Association for Computational Linguistics. DOI: 10.18653/v1/K18-2020.
- Tufiş, Dan & Verginica Barbu Mititelu. 2014. The lexical ontology for Romanian. In Nuria Gala, Reinhard Rapp & Nuria Bel-Enguix (eds.), *Language Production, Cognition, and the Lexicon*, 491–504. Cham: Springer.
- Tufiş, Dan, Dan Cristea & Sophia Stamou. 2004. BalkaNet: Aims, methods, results and perspectives. A general overview. *Romanian Journal of Information Science and Technology* 7(1-2). 9–43.
- Vietri, Simonetta. 2014a. *Idiomatic constructions in Italian: A lexicon-grammar approach*. Amsterdam/Philadelphia: John Benjamins.
- Vietri, Simonetta. 2014b. The lexicon-grammar of Italian idioms. In Jorge Baptista, Pushpak Bhattacharyya, Christiane Fellbaum, Mikel Forcada, Chu-Ren Huang, Svetla Koeva, Cvetana Krstev & Éric Laporte (eds.), *Proceedings of the workshop on lexical and grammatical resources for language processing (LG-LP 2014)*, 137–146. Dublin, Ireland: Association for Computational Linguistics & Dublin City University. DOI: 10.3115/v1/W14-5817.
- Villavicencio, Aline, Timothy Baldwin & Benjamin Waldron. 2004a. A multilingual database of idioms. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC'04)*, 1127–1130. Lisbon. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/760.pdf>.
- Villavicencio, Aline, Ann Copestake, Benjamin Waldron & Fabre Lambeau. 2004b. Lexical encoding of MWEs. In Takaaki Tanaka, Aline Villavicencio, Francis Bond & Anna Korhonen (eds.), *Proceedings of the second ACL workshop on multiword expressions: Integrating processing*, 80–87. Barcelona: ACL.

Chapter 4

Representation of multiword expressions in the Bulgarian integrated lexicon for language technology

👤 Petya Osenova^a & 👤 Kiril Simov^a

^aInstitute of Information and Communication Technologies, Bulgarian Academy of Sciences

The chapter introduces a representation model of multiword expressions from the perspective of integrated lexicons for Bulgarian. The lexicons considered are an inflectional one, a valency one, and a wordnet. We created a joint representation entry that incorporates morphology, valency potential and lexical semantics through synonym sets. The selected mechanism for displaying all the information is catenabased since the catena allows for better modeling of idiosyncratic elements and is tree-based. Also, a general typology of multiword expressions is proposed that focuses on fixedness and (dis)continuity. We believe that providing a unified representation of multiword expressions and common lexica would improve the performance of the various natural language processing applications.

1 Introduction

This paper is based on our previous investigations on multiword expressions (MWEs) for Bulgarian (Simov & Osenova 2015a, Laskova et al. 2019). This previous research was motivated by the investigation of the most adequate representations of MWEs in treebanks, in syntax-aware lexicons like the valency ones and in lexical bases like wordnets.

Having already developed a number of language resources for Bulgarian, our current goal is to integrate them in such a way that they would allow a joint



approach to several NLP (natural language processing) tasks, including end-to-end training of neural network models.

In order to achieve this goal, we have already integrated the Bulgarian treebank (BTB) with sense annotations from the Bulgarian wordnet (BTB-WN), Bulgarian DBpedia, Bulgarian Wikipedia, Bulgarian Valency Lexicon, and a newly created small FrameNet-oriented lexicon for event annotation in the area of Digital Humanities. With respect to the integrated lexical and text resources, one of the problems is the common representation of the lemmas in the various types of lexicons, especially the representation of MWEs. Thus, one of the important requirements is that lemmas have a common representation in both – the annotated corpora and the integrated lexical resources. However, other issues appear here: what the lemma of a MWE is; how to present the syntactic potential in a lexical database including the points of flexibility and external participants; and how to map the lexical representation to the one in a corpus.

In this paper, we focus on the representation of MWEs in the framework of integrated lexical resources. In relation to that our contributions are as follows:

1. introducing the structure of the MWE lexical entry;
2. tuning the catena-based formalization to the complex structure of integrated linguistic information;
3. modeling the complexity of the entry with respect to discontinuity and fixedness.

The paper is organized as follows: in §2 related work is discussed. §3 introduces the background of our model. §4 introduces the formal definition of catena. §5 presents a model of the lexical entry. §6 suggests analyses of the specific MWE types. §7 concludes the paper.

2 Related work

The representation of MWEs in lexicons with a view to their adequate annotation in corpora has been a hot topic for quite some time. For example, Lichte et al. (2019) discuss various approaches to lexical encoding of MWEs with respect to the NLP tasks. The authors favor flexible formats like PATRII and XMG over the fixed encoding formats of a Dutch Electronic Lexicon of Multiword Expressions (Grégoire 2010), and a Polish Valency Lexicon (Przepiórkowski et al. 2014). Our current approach is somewhere between the fixed and flexible encodings. On the one hand, it uses property name sets where the main morphosyntactic, syntactic,

and semantic characteristics of the MWE are given. At the same time, the notion of catena is used, which introduces a graph representation and thus falls into the tree-based approaches to MWEs. In this way, the catena ensures the flexibility of the encoding with respect to potential discontinuity or other specifics. Our approach is head-based rather than construction-based.

Dyvik et al. (2019) present the encoding of MWEs in the resource grammar NorGram which is based on the Lexical-Functional Grammar (LFG) framework. There the fixed MWEs are treated as words. For the flexible MWEs another approach is taken – namely, following the grammar apparatus of LFG, the components are presented through selection frames with a subcategorization in case of verbs and complements, and with equations for the other lexically restricted dependants – all these with their specifics. In this paper, the approach is lexico-syntactic since the representation of the MWEs combines both – the morphosyntactic and lexical specifics. Thus, through the theory mechanisms, the balance between grammar and lexicon is pertained. Our approach aims to ensure exactly such a dynamic relation between a lexicon and a grammar without the availability of a well-developed computational grammar.

Masini (2019) introduces three criteria for classifying MWEs: “(i) formal properties (degree of internal cohesion or fixity), (ii) idiomatic status [...], and (iii) function, or a combination of these”. In our proposed approach we focus mainly on (i) under which we also include (ii). Then we are more interested in the challenges when modeling word order than in the function of the MWE per se (see §5).

There are attempts for MWE representation in dictionaries and databases for both – humans and machines, i.e. reflecting multipurpose and multilevel aspects. For example, Vondříčka (2019) uses slots for the syntagmatic information and fillers for the paradigmatic one in the entry. The author relies on the tree representation in dependency and constituency formats with the accompanying challenges. The problems come from the notion of the word and ways of spelling as well as from the not straightforward modeling of the internal elements in a MWE. In Skoumalová et al. (2024 [this volume]) the linking is described of the lexical entries in a MWE lexicon for Czech with their natural occurrences in a corpus. The relation between the lexicon and the corpus has been ensured in both directions. We aim at such an integrated resource and workflow. However, at the moment we provide a link of a MWE to its corpus occurrence only through the headwords of MWEs.

In Lion-Bouton et al. (2023) the authors propose an approach according to which the MWE identification tools consult lexicons. For this purpose, a survey has been performed on quantitative evaluation of some MWE lexicon formalisms based on the notion of observational adequacy. The suggested approach based

on a generalisation of the concept of a Coarse Syntactic Structure proves to be competitive with lexicons based on a sequential representation of MWEs. Our approach is also graph/tree-based but we aim to accommodate as much information as possible in the same representation – lexical from wordnets, valency from valency dictionaries, knowledge-based from Wikipedia, etc.

Zampieri et al. (2019) show the impact of the MWE representation in the input pre-processed data as well as in two types of word embeddings (word2vec and FastText) for the task of MWE identification. They conclude that the lemma plays a positive role for all considered languages – Basque, French, and Polish. For us the most interesting part in relation to our work is the fact that the richer the information for a morphologically rich language, the better the results. We also try to represent as much integrated information about a MWE as possible.

Schneider et al. (2014) report on the annotation of MWEs in a social web corpus. They use an annotation scheme that respects the following aspects: heterogeneity (where the annotated MWEs are not restricted by syntactic construction); shallow but gappy grouping (MWEs viewed as simple groupings of tokens, which need not be contiguous in the sentence); and expression strength (where the most idiomatic MWEs are distinguished from and can belong to weaker collocations). For our work the most important focus (along the others) is the modeling of gapping, i.e. discontinuity. Authors indicate that 15% of MWEs contain at least one gap. We have to take into account that this fact is given for English as a language with a rather fixed word order. In languages like Bulgarian that have a relatively free word order, discontinuity is expected to be much higher. For that reason we are trying to find a way to model the predicted points of discontinuity within the lexical entry.

In Leseva et al. (2024 [this volume]) an elaborate bilingual model of MWEs representation is described for Bulgarian and Romanian in a uniform way. Wordnets for the two languages have been used for linking the bilingual lexicons. The focus is put on the verbal MWEs where the relations from the Universal Dependencies (UD) have been used. We also use a wordnet for Bulgarian (BTB-WN) as a linking module and UD as modeling relations within MWEs.

In the PARSEME initiative verbal MWE (VMWE) annotations, both continuous and discontinuous groups are considered (Savary et al. 2018). The annotation strategy includes the lexicalized elements, not their variations. It views the representation as a syntactic tree. However, the scheme describes also the properties for each type and provides specialized guides for each participating language, including Bulgarian. In addition to the two universal VMWE categories (light verb constructions with two subtypes and verbal idioms), our language has inherently reflexive verbs (IRV) but not verb-particle constructions (VPC). Since

our task here is to show how we represent all the main types of MWEs, we focus on the variety and complexity of their modeling.

3 Background

Our work on MWEs up to now has been centred around the notion of catena. Catena (chain) was initially introduced in O’Grady (1998) as a mechanism for representing the syntactic structure of idioms. He showed that for this task a definition of syntactic patterns was needed that does not coincide with constituents. He defined the catena in the following way: “The words A, B, and C (order irrelevant) form a chain if and only if A immediately dominates B and C, or if and only if A immediately dominates B and B immediately dominates C”. Some examples of catena from a dependency syntactic tree are presented in Figure 1. In our work here we convert MWEs into a representation previously defined in Simov & Osenova (2014) and in Simov & Osenova (2015b) in which the catena is depicted as a dependency tree fragment with appropriate grammatical and semantic information. The variations of the MWEs are represented through underspecification of the corresponding features, including valency frames and non-canonical basic form.

The lexical entry uses the following format: a lexicon catena (LC), semantics (SM) and valency (Frame). The lexicon catena for the MWEs is stored in its basic form. The realisation of the catena in a sentence has to obey the rules of the grammar. In this way the possible word order is managed. The semantics of a lexical entry specifies the list of elementary predicates contributed by the lexical item. When the MWE allows for some modification (including adjunction) of its elements, i.e. modifiers of a noun, the lexical entry in the lexicon needs to specify the role of these modifiers. Some first ideas in these lines are represented in the above cited works and also in Laskova et al. (2019).

We aim at an integrated and relatively flexible representation of MWE types in lexicons and their projections in corpora. We are aware that this task is not trivial and will take time. Our proposal builds on our previous modelling. Here we discuss an extended lexical entry model in order to incorporate as much linguistic information as possible. In our previous publications we already assumed that each lemma in the lexicon is represented as a catena (even when it is not a MWE). This assumption allows us to represent information in relation to analytical verb forms, to the order of the component words in the MWEs, to their morphosyntactic variations, to their syntactic and semantic behaviour, to the etymological information in cases when peculiarities of MWEs have diachronic origin. For example, in the Bulgarian expression (bg) *добър вечер* *dobar vecher* (lit. ‘good-SG.M

evening-SG.F') 'good evening', 'good' is masculine and 'evening' is feminine. The surface agreement is violated because the noun 'evening' changed its gender in contemporary language to feminine.

The model of the Valency lexicon follows our insights from the catena representation of MWEs. Such an approach allows us to introduce the integration of the necessary world knowledge to the frame elements, especially the interaction among the types of participants within a given event. Needless to say, this kind of information is not always fully compositional and the boundaries between compositional and non-compositional are not always clear. Thus, we think that the same lexicon model can be applied to the continuum from compositionality to non-compositionality in a valency-aware dictionary. We imagine that this effort will not be deterministic but incremental, since MWEs show idiosyncrasies all the time across genres, alternations, figurative meanings, etc.

Our main contribution in this paper is the structure of the lexical entry in an integrated lexicon by means of the catena notion. In the integrated resource we have included the following distinct lexicons:

Inflectional lexicon of Bulgarian (ILB): Each lemma is connected to its inflectional paradigm;

BTB Bulgarian WordNet (BTB-WN): A Bulgarian WordNet which arranges synonym sets around identical meanings. The lexical entry in BTB-WN is called *SYNSET (Synonym Set)*;

Bulgarian Valency Lexicon (BVL): Complex representation of the core participants of a given event (in general sense) represented by a verb in its meaning.

The main decision we took was about the mechanism for integrating lexical entries from these three lexicons: ILB, BVL and BTB-WN. First, the initial representation of the original lexical entries is introduced. Note that we omit details that are not important for this paper. Such details, for example, include the interaction between the lexical and semantic relations in the BTB-WN.

The lexical entry of ILB includes the following main elements: *Lemma*, *Part of speech*, and *Paradigm*. The lemma is the abstract representation of the lexical entry. Each part of speech is one of the ten common parts of speech in Bulgarian (noun, adjective, numeral, adverb, pronoun, verb, preposition, conjunction, particle, interjection). For a detailed description of Bulgarian see Osenova (2010). The paradigm is a list of all the synthetic word forms related to the lemma. Bulgarian is an analytical and inflectional language. It has a rich inflectional morphology,

but listing all the members of the synthetic part of the verb paradigm is still feasible, because the largest paradigm contains only 52 word forms. Each word form corresponds to a given set of grammatical features. Some word forms are analytical like part of the Bulgarian tenses. For example, the verb (bg) *чета cheta* (lit. read-1SG.PRS) ‘I read’ forms a future tense, second person, singular as follows: (bg) *ще четеш shte chetesh* (lit. read-2SG.FUT) ‘you will read’. Such analytical word forms are formed by patterns (rules) which we consider as a part of the lexicon. They are represented using the same mechanism as the rest of the lexicon.

The Lexical entry of BTB-WN includes the following main elements: *Definition*, *Set of synonyms*, *Examples*. Each definition in BTB-WN provides a description of the meaning in Bulgarian. The set of synonyms is represented via a set of lemmas sharing the meaning of the synset. Each lemma is connected to a paradigm and a part of speech. Each example consists of one or more sentences in which the corresponding meaning is exemplified. Each example in a synset is also linked to its lemma. We usually include only one sentence, but if one sentence is not enough to disambiguate between the different meanings of the lemma, then more sentences are included. Also, the example is linked to the source from where it is taken. In this way, if necessary, we could extract more data. The current version of BTB-WN contains 53217 lemmas of which 7868 are MWEs (14.78%).

The lexical entry of BVL includes the following main elements: *Lemma*, *Definition*, *Valency frame*, and *Examples*. The lemma is the verb lemma for the lexical entry. Each definition represents a meaning of the lemma. The definition is the same as in the wordnet. The valency frame introduces a generalised representation of the core participants of the event denoted by the meaning and the syntactic behaviour of the lemma as well as by the core participants. The current version of BVL contains 6869 lemmas 1674 of which are MWEs (24.37%).

In order to integrate the lexical entries of the three lexicons we followed the following procedure:

- *Achieving a uniform representation of lemmas*. Since the three lexicons were constructed in different periods and on the basis of different machine readable sources, the lemmas of the same word could have had different representations. This holds especially for the ILB – the lexicon whose first version was created earlier.
- *Mapping of the meanings*. We have ensured that the meaning in BTB-WN and BVL are the same for the respective verbs. In this way, the verb lemmas and meanings in BTB-WN and BVL have been unified.

- *Modification of the paradigm.* Since the paradigm sometimes depends on the meaning of the lemma, the paradigm inherited from ILB had to be modified in a number of cases. For example, some nouns in some meanings are only pluralia tantum.

Thus, the lexical entry of the integrated lexicon consists of two elements: (Definition and Set of synonyms). The information about the paradigm, valency frames and examples is represented within the entry of each lemma. The record for each lemma contains also a link to its paradigm; one or more valency frames; a set of examples; and other lemma dependant classifications.

Each lemma is converted into its syntactic representation as a catena (see next section). When the lemma is a single word, the conversion to a catena is trivial. At the same time, the complexity of MWEs requires more attention to the construction of the appropriate representation. For more details see next sections. In addition to the synthetic forms, the verb paradigm contains also the analytical ones. We consider them as a special class of MWEs. The patterns for the analytical forms are represented as an addition to the main lexicon. In the lexical entry only a link to the corresponding set of patterns is given.

4 Formal definition of catena

In this section we define the formal presentation of the catena as it is used in syntax and in the lexicon. Here we follow the definition of catena provided by O’Grady (1998) and Groß (2010): a CATENA is a word or a combination of words directly connected in the dominance dimension. In reality, this definition of catena for dependency trees is equivalent to a subtree definition. Figure 1 depicts a complete dependency tree and some of its catenae. Notice that the complete tree is also a catena itself. With “root_C” we mark the root of the catena. It might be the same as the root of the complete tree, but also different as in the cases of “John” and “apple”. Following Osborne et al. (2012) we prefer to use the notion of catena to that of dependency subtree or treelet. We aim to utilize the notion of catena for several purposes: representation of words and MWEs in the lexicon, their realization in the actual trees that present the sentence analysis, as well as for the representation of the derivational structure of compounds in the lexicon.

In order to model the variety of phenomena and characteristics encoded in a dependency grammar we extend the catena with partial arc and node labels. We follow the approach taken in CoNLL shared tasks on dependency parsing (Buchholz & Marsi 2006) representing for each node its word form, lemma, part of speech, extended part of speech, grammatical features (and later – semantics). This provides a flexible mechanism for expressing the combinatorial potential of

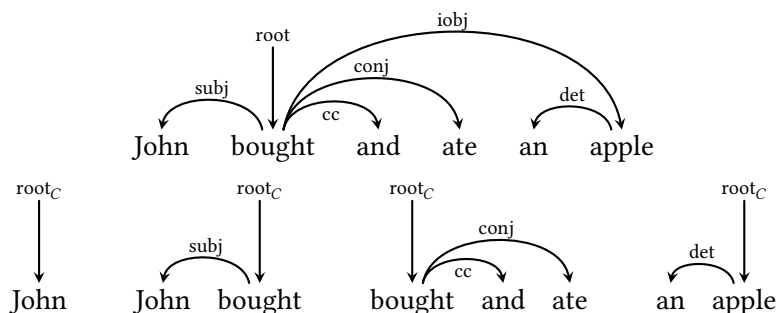


Figure 1: A complete dependency tree and some of its catenae. The complete list of catenae of the complete tree is too large to be presented here.

lexical items. In the following definition all grammatical features are represented as part-of-speech (POS) tags.¹

Let us have the sets: LA – a set of POS tags,² LE – a set of lemmas, WF – a set of word forms, and a set of dependency tags D ($root \in D$). Let us have a sentence $x = w_1, \dots, w_n$. A TAGGED DEPENDENCY TREE is a directed tree $T = (V, A, \pi, \lambda, \omega, \delta)$ where:

1. $V = \{0, 1, \dots, n\}$ is an ordered set of nodes that corresponds to an enumeration of the words in the sentence (the root of the tree has an index 0);
2. $A \subseteq V \times V$ is a set of arcs. For each node i , $1 \leq i \leq n$, there is exactly one arc in A : $\langle i, j \rangle \in A$, $0 \leq j \leq n$, $i \neq j$. There is exactly one arc $\langle i, 0 \rangle \in A$;
3. $\pi : V - \{0\} \rightarrow LA$ is a total labelling function from nodes to POS tags.³ π is not defined for the root;
4. $\lambda : V - \{0\} \rightarrow LE$ is a total labelling function from nodes to lemmas. λ is not defined for the root;
5. $\omega : V - \{0\} \rightarrow WF$ is a total labelling function from nodes to word forms. ω is not defined for the root;

¹In fact, our tagset encodes all the morphosyntactic tags related to each part-of-speech, but here we use the notion of POS tag as a more common term. The tagset is described here: <http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR03.pdf>.

²In the formal definitions here we use tags as entities, but in practice they are sets of grammatical features like values for gender, number, etc.

³In case we are interested in part of the grammatical features encoded in a POS tag we could consider π as a set of different mappings for the different grammatical features. It is easy to extend the definition in this respect, but we do not do this here.

6. $\delta : A \rightarrow D$ is a total labelling function for arcs corresponding to the dependency label. Only the arc $\langle i, 0 \rangle$ is mapped to the label *root*;
7. 0 is the root of the tree.

We will hereafter refer to this structure as a parse tree for the sentence x . Node 0 does not correspond to a word form in the sentence, but plays the role of a root of the tree.

Let $T = (V, A, \pi, \lambda, \omega, \delta)$ be a tagged dependency tree. A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ is called DEPENDENCY CATENA OF T if and only if there exists a mapping $\psi : V_G \rightarrow V^4$ such that:

1. $A_G \subseteq A$, the set of arcs of G ;
2. $\pi_G \subseteq \pi$ is a partial labelling function from nodes of G to POS tags;
3. $\lambda_G \subseteq \lambda$ is a partial labelling function from nodes of G to lemmas;
4. $\omega_G \subseteq \omega$ is a partial labelling function from nodes of G to word forms;
5. $\delta_G \subseteq \delta$ is a partial labelling function for arcs of G to dependency labels.

A directed tree $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ is a DEPENDENCY CATENA if and only if there exists a dependency tree T such that G is a dependency catena of T . We mark the root catena with $root_C$ arc in graphical representation.

The partial functions for assigning POS tags, dependency labels, word forms and lemmas allow us to construct arbitrary abstractions over the structure of a catena. Thus, the catena could be underspecified for some of the node labels, like grammatical features, lemmas and also some dependency labels. In this way the catena could be a dependency catena of dependency trees which differ with respect to labels of different kinds. Thus, catenae are a good choice for encoding variability of lexical representation of MWEs.

Thus mapping ψ parameterizes the catena with respect to different dependency trees. Using the mapping, there is a possibility to realize different word orders of the catena nodes, for instance. The omission of node 0 from the range of the mapping ψ excludes the external root of the tagged dependency tree from each catena. The catena could be a word or an arbitrary subtree.

⁴This mapping allows for embedding of G in different tagged dependency trees and thus different word order realizations of the catena nodes (corresponding to word forms in T). The mapping ψ is specific for G and T . It allows also the image of G in T not to be a subtree of T , but several subtrees of T . A special case is discussed below – partition and extension operations.

We call the mapping of a catena into a given dependency tree the REALIZATION OF THE CATENA IN THE TREE. We consider the realization of the catena as a fully specified subtree including all node and arc labels. For example, the catena for “to spill the beans” will allow for any realization of the verb form like in: “they spilled the beans” and “he spills the beans”. Thus, the catena in the lexicon will be underspecified with respect to the grammatical features and word forms for the verb.

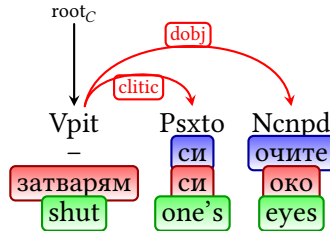
This underspecified catena will be called a LEXICON CATENA (LC), because it will be stored in the lexical entries. Figure 2 depicts two realizations (with different word orders) of the catena for the idiom (bg) затварям си очите *zatvaryam si ochite* (lit. shut-1SG.PRS REFL eyes-DEF) ‘I ignore the facts’. The upper part of the image represents the lexicon catena for the idiom. It determines the fixed elements of the catena: the arcs, their labels, the nodes and their labels: extended part of speech (first row), word forms (second row), lemmas (third row), and gloss in English (fourth row).⁵ The dash (–) in the word form row means that the word form is not defined for the verbal node. In this way the word form could be different in the different realization of the catena. Also, the POS tag in the catena is underspecified with respect to features of the different word forms. In the two realizations, the verbal forms received their specific tags. Also, fixed elements of the catena are represented as in the image of the catena. The word order in the two realizations is different. Thus, catenae with different underspecified elements define different levels of freedom in the realization of the MWEs.

Let G_1 and G_2 be two catenae. A COMPOSITION of G_1 and G_2 is a catena G_c , such that

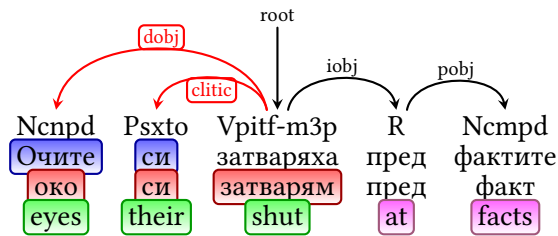
1. the catenae G_1 and G_2 are realized in catena G_c ,
2. each node in catena G_c is an image of a node from G_1 or G_2 , or both,
3. the root of catena G_c is an image of the root of catena G_1 ,
4. if a node i in catena G_c is an image of node i_1 in catena G_1 and i_2 in G_2 , then all the information assigned to these nodes is compatible and fully represented in the node i ,
5. if an arc $\langle i, j \rangle$ in catena G_c is an image of arc $\langle i_1, j_1 \rangle$ in catena G_1 and $\langle i_2, j_2 \rangle$ in G_2 , then the label of $\langle i, j \rangle$ if it exists, has to be compatible with the labels of the arc $\langle i_1, j_1 \rangle$ in G_1 and $\langle i_2, j_2 \rangle$ in G_2 .

⁵In the next examples we present only the important information, thus, some of these rows will be missing. In other cases new rows will be used to represent additional information.

Lexicon catena:



Realization 1: (bg) Очите си затваряха пред фактите *Ochite si zatvaryaha pred faktite* (lit. eyes-DEF REFL shut-3PL.PST.PROG at facts-DEF) ‘They ignored the facts’:



Realization 2: (bg) Иван си затваряше очите *Ivan si zatvaryashe ochite* (lit. Ivan-SG REFL shut-3SG.PST.PROG eyes-DEF) ‘Ivan ignored the facts’:

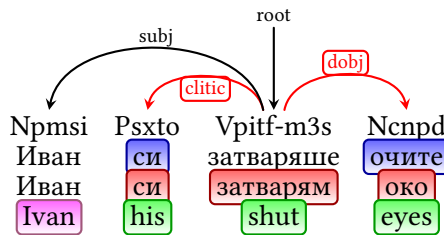


Figure 2: Two realizations of the lexicon catena for the idiom (bg) затварям си очите *zatvaryam si ochite* (lit. shut-1SG.PRS REFL eyes-DEF) ‘I ignore the facts’.

The lemma information for two nodes i_1 in G_1 and i_2 in G_2 is compatible if at least one of the nodes does not have an assigned lemma, or if both nodes have the same assigned lemma. It is similar for word forms. For POS tags the compatibility is defined as a tag representation that contains the information of tags defined for both nodes. For example, if we have partial POS tag specifications ‘Vpit’ and ‘Vp-m2s’, the compatible specification is ‘Vpit-m2s’. The arc labels are compatible if and only if they are the same, or at least one of them is not defined. If for both arcs the labels are not defined, then the label for the image arc is also not defined. Similar definitions could be stated for any other information added to the nodes and arcs such as semantic information, etc.

Using the composition operation we could realize the selectional restrictions of a given lexical unit with respect to a catena in a sentence.

For example, let us assume that the verb ‘to read’ requires the subject to be a human and the object to be an information object. In Figure 3 we present how the catena for ‘I read’ is combined with the catena ‘a book’ in order to form the catena ‘I read a book’. The figure represents only the level of word forms and a level of semantics (specified only for the node on which the composition is performed). The catena for ‘I read ...’ specifies that the unknown direct object has the semantics of an *Information Object (InfObj)*. The catena for ‘a book’ represents the fact that the book is an Information Object. Thus the two catenae could be composed on the two nodes marked as InfObj. The result is represented at the bottom of Figure 3.⁶

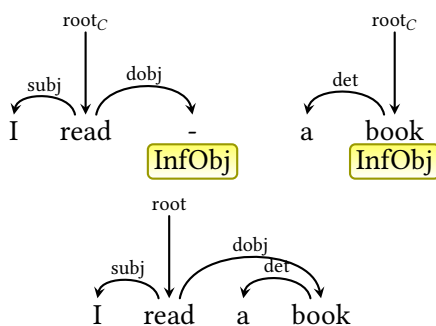


Figure 3: Composition of catenae.

⁶In this representation many details like lemmas and grammatical features are not presented because they are not important for the example.

Some MWEs require more complex operations over catenae. Such a class of MWEs are idioms with a lexicalized subject, such as “the devil is in the details”; the realizations of catenae from the lexicon into dependency trees are often accompanied by intervening material – see the discussion in Osborne et al. (2012). For example, the above-mentioned idiom allows realizations such as: “the devil will be in the details”, “the devil seems to be in the details”, etc. Thus we need to modify the internal structure of the lexicon catena.

Our insight, supported by the examples, is that the intervening material forms a catena of a certain type. Such a type of catena will be called an **AUXILIARY CATENA**⁷ in this paper, although it could be of different kinds (auxiliary, modal, control, etc.), depending on the verb forms. In order to implement this idea we need some additional notions.

Let $G = (V_G, A_G, \pi_G, \lambda_G, \omega_G, \delta_G)$ be a catena and $k \in V_G$ and m is integer and $m > 1$, then G_1, G_2, \dots, G_l is a partition of G on node k if and only if:

1. each G_i for $1 \leq i \leq m$ is a catena which is a subtree of G ;
2. one or more subcatenae G_i for each $1 \leq i \leq m$ have k as a root node;
3. the only common node for all subcatenae G_i is k ;
4. the mappings $\pi_{G_i}, \lambda_{G_i}, \omega_{G_i}, \delta_{G_i}$ are the same as for the whole catena G , except for the node k where the mappings $\pi_{G_i}, \lambda_{G_i}, \omega_{G_i}$ could be partial with respect to the original mappings.

An example of the operation **partition** of *the devil is in the details* is given in Figure 4.

After the partition of the catena, we need a mechanism to connect the different catenae of the partition with the auxiliary catena.

Let G be a catena and for $n \in V_G, G_1, G_2, \dots, G_n$ be a partition of G and G_a be an auxiliary catena. An **EXTENSION** of G on partition G_1, G_2, \dots, G_n with catena G_a is a catena G_e such that each catena G_1, G_2, \dots, G_n and the auxiliary catena G_a are realized in G_e in such a way that the node n_i in G_i (corresponding to the original node n) is mapped to a node in G_e to which a node of G_a is mapped. Each node in G_e is an image of a node from G_1, G_2, \dots, G_n or G_a .

An example of the operation **extension** is presented in Figure 5.⁸

⁷Under auxiliary catena we assume a catena that is part of the verbal complex (i.e. an analytical tense of a verb, where elements such as clitics can be inserted between components) and contains nodes for the auxiliary verbs. In the grammars for the different languages different kinds of catena could be defined on the basis of their role in the grammar. In this respect, the definition of extension here is restricted to the verbal complex, but could be easily adapted for other cases when necessary.

⁸Note that there are alternative analyses in which the auxiliary verb is not a head of the sentence, but a dependent of the copula.

4 Representation of multiword expressions in Bulgarian

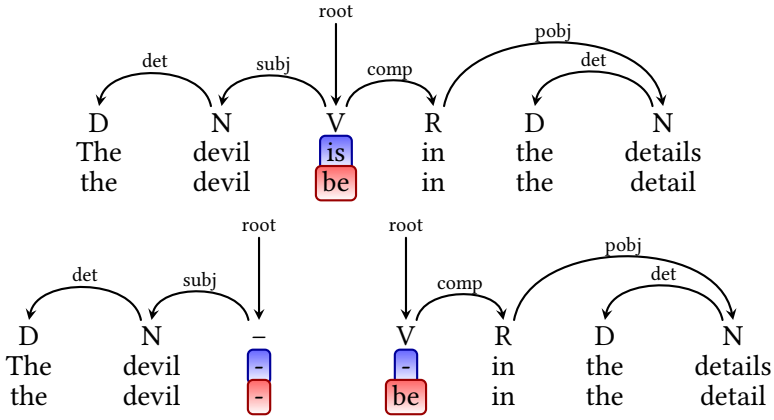


Figure 4: Partition of the catena for “the devil is in the details”.

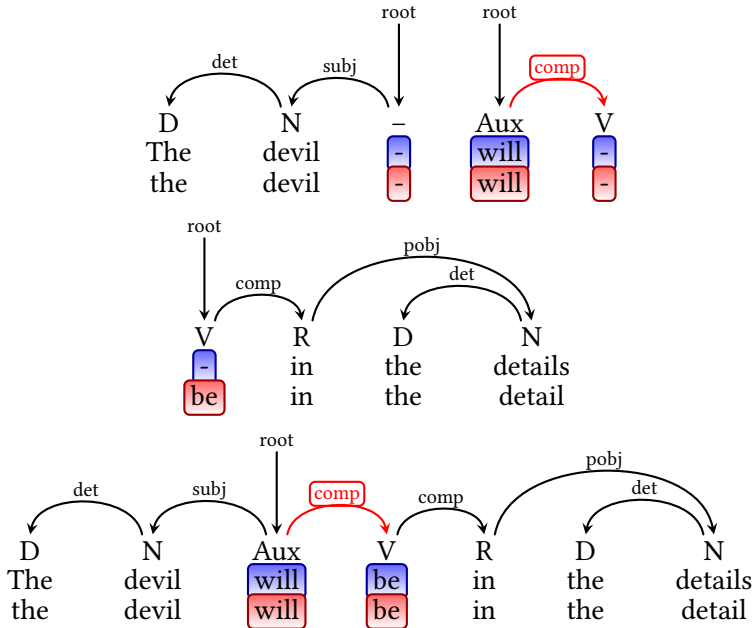


Figure 5: Extension.

Two catenae G_1 and G_2 could have the same set of realizations. In this case, we will say that G_1 and G_2 are EQUIVALENT. Representing the nodes via paths in the dependency tree from root to the corresponding node and imposing a linear order over this representation of nodes facilitates the selection of a unique representative of each equivalent class of catenae. Thus, in the rest of the paper we assume that each catena is representative of its class of equivalence. This representation of a catena will be called CANONICAL FORM.

5 A model of a lexical entry

In this section we use the notion of catena already introduced in Section 4, to define in greater detail the structure of a lexical entry as presented above. Through the operations of *composition*, *partition* and *extension* it becomes possible to compose the different parts of this structure and thus manage the actual realization of the lexical items in text. In this paper we represent the syntactic information in terms of the dependency grammar, but it can be done in a similar way within phrase-based grammars.

For each node in a catena or a dependency tree we present the following information: POS, Grammatical Features, Word Form, Lemma, Node identifier (the position of a word form in a catena or a sentence). Each piece of information is depicted in the node representation at a different row.

In Figure 6 a model of the lexical entry is presented. Each lexical entry for a synset includes (minimally): *Synset* which defines the synset information and *SynsetID* which identifies the synset in a unique way; *Definition* which expresses the content of the meaning of the synset; *Lemma list* which contains the representation of each lemma that shares the meaning of the synset. Each lemma is represented by the following elements: *LemmaID* which introduces the lemma in a unique way in the whole lexicon; *Basic Form* is a selected word form from the paradigm of the lemma; *Paradigm* is a list of pairs consisting of a word form, represented as a catena, and a tag, encoding the grammatical features of the word form. Each word form is a catena; *Valency Frame* represents the selectional restrictions of the lemma. The valency frame is represented as a catena. *Examples* is a list of example sentences or short texts. The realization of a lemma in a text requires the selection of the appropriate word form from the paradigm, represented as a Word Form Catena (WFC), composed with the Valency Frame Catena (VFC).

In Figure 7 we give an example lexical entry for the verb (bg) бягам *byagam* (lit. run-1SG.PRS) ‘to run’. The most important information is presented in the following sections: *Paradigm*, where we could see two catenae for *present tense, first person, singular*, and *present tense, second person, singular*, and in *Valency frame* (V. Frame) where a catena for the valency restrictions is given.

4 Representation of multiword expressions in Bulgarian

Synset: <i>Example entry</i> Synset ID: <i>SynsetID</i>													
Definition: <i>Text of the definition</i>													
Lemma list:	...												
	<table border="1"> <tr> <td>LemmaID:</td> <td><i>Lemma-ID1</i></td> </tr> <tr> <td>Basic Form:</td> <td><i>BasicForm-Lemma-ID1</i></td> </tr> <tr> <td>Paradigm:</td> <td> WordForm₁₁ : GrammaticalTag₁₁ WordForm₁₂ : GrammaticalTag₁₂ ... WordForm_{1n} : GrammaticalTag_{1n} </td> </tr> <tr> <td>Valency Frame:</td> <td><i>Valency Frame Description</i></td> </tr> <tr> <td>Examples:</td> <td><i>List of examples for this lemma</i></td> </tr> <tr> <td>Analytical Class:</td> <td><i>Pattern Class</i></td> </tr> </table>	LemmaID:	<i>Lemma-ID1</i>	Basic Form:	<i>BasicForm-Lemma-ID1</i>	Paradigm:	WordForm ₁₁ : GrammaticalTag ₁₁ WordForm ₁₂ : GrammaticalTag ₁₂ ... WordForm _{1n} : GrammaticalTag _{1n}	Valency Frame:	<i>Valency Frame Description</i>	Examples:	<i>List of examples for this lemma</i>	Analytical Class:	<i>Pattern Class</i>
	LemmaID:	<i>Lemma-ID1</i>											
	Basic Form:	<i>BasicForm-Lemma-ID1</i>											
	Paradigm:	WordForm ₁₁ : GrammaticalTag ₁₁ WordForm ₁₂ : GrammaticalTag ₁₂ ... WordForm _{1n} : GrammaticalTag _{1n}											
	Valency Frame:	<i>Valency Frame Description</i>											
	Examples:	<i>List of examples for this lemma</i>											
	Analytical Class:	<i>Pattern Class</i>											
	...												
	<table border="1"> <tr> <td>LemmaID:</td> <td><i>Lemma-IDK</i></td> </tr> <tr> <td>Basic Form:</td> <td><i>BasicForm-Lemma-IDK</i></td> </tr> <tr> <td>Paradigm:</td> <td> WordForm_{K1} : GrammaticalTag_{K1} WordForm_{K2} : GrammaticalTag_{K2} ... WordForm_{Kn} : GrammaticalTag_{Kn} </td> </tr> <tr> <td>Valency Frame:</td> <td><i>Valency Frame Description</i></td> </tr> <tr> <td>Examples:</td> <td><i>List of examples for this lemma</i></td> </tr> <tr> <td>Analytical Class:</td> <td><i>Pattern Class</i></td> </tr> </table>	LemmaID:	<i>Lemma-IDK</i>	Basic Form:	<i>BasicForm-Lemma-IDK</i>	Paradigm:	WordForm _{K1} : GrammaticalTag _{K1} WordForm _{K2} : GrammaticalTag _{K2} ... WordForm _{Kn} : GrammaticalTag _{Kn}	Valency Frame:	<i>Valency Frame Description</i>	Examples:	<i>List of examples for this lemma</i>	Analytical Class:	<i>Pattern Class</i>
	LemmaID:	<i>Lemma-IDK</i>											
	Basic Form:	<i>BasicForm-Lemma-IDK</i>											
	Paradigm:	WordForm _{K1} : GrammaticalTag _{K1} WordForm _{K2} : GrammaticalTag _{K2} ... WordForm _{Kn} : GrammaticalTag _{Kn}											
	Valency Frame:	<i>Valency Frame Description</i>											
Examples:	<i>List of examples for this lemma</i>												
Analytical Class:	<i>Pattern Class</i>												

Figure 6: Lexical entry model.

The information related to the nodes in the catena is represented on different layers as follows: the bottom row contains the names of the corresponding nodes: CNo1, CNo2, etc. (in many examples in the paper this information is not presented, because it is redundant to a certain extent); the next row up contains the translation of the word form in English; the next two rows up are for the lemma of the node and for the word form. If the word form row contains “-” then the node is underspecified for a word form and it is determined by another catena during the composition operation. The last two rows up represent the grammatical features for the corresponding word forms. The first row contains information for each word form in its own lexical entry. The second row (the top one) contains grammatical information for the node when it is realized in the complete word form. When the word form is a single word, then the value in the two rows coincides. The difference could appear when in MWEs (including

Synset: бягам от отговорност Synset ID: SID-003592					
Definition: Отбягвам да поема отговорност					
Lemma list:	<table border="1"> <tr> <td>LemmaID:</td> <td><i>btbwn-041000447-v</i></td> </tr> <tr> <td>B. Form:</td> <td><i>бягам</i></td> </tr> </table>	LemmaID:	<i>btbwn-041000447-v</i>	B. Form:	<i>бягам</i>
	LemmaID:	<i>btbwn-041000447-v</i>			
	B. Form:	<i>бягам</i>			
	Paradigm:				
	V. Frame:				
Examples:	<i>List of examples for this lemma</i>				
Analytical Class:	<i>PatternClassVp</i>				

Figure 7: Lexical entry for the verb (bg) бягам (от отговорност) *byagam (ot otgovornost)* (lit. run-1SG.PRS (from responsibility-SG.F)) ‘to run away from one’s responsibility’.

analytical forms) some of the grammatical features are modified. In the example above, the word form for future tense is composed of the auxiliary particle (bg) *ще* *ste* (lit. will-FUT) ‘will’ and the verb form for *present tense, second person, singular*. The whole word form is in future tense. In the example, the morphosyntactic tag *Vpiif-r2s* (tag for present tense) becomes *Vpiif-f2s* (tag for future tense in an analytical verb form). In the text realization we perform composition of one catena from the paradigm and the catena from the valency frame. Thus, the result from this operation between the analytical word given above and the valency catena results in the following catena – see Figure 8.

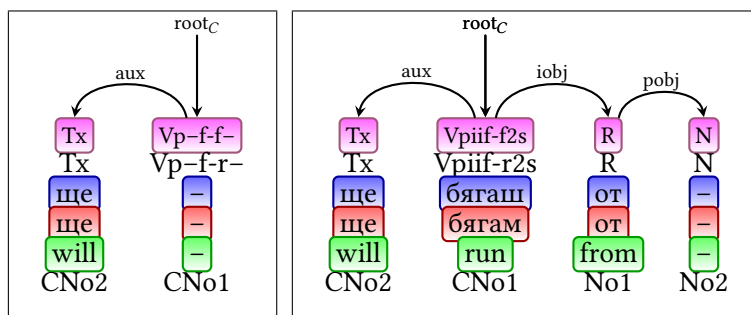


Figure 8: On the left, the auxiliary catena for future tense is given. As can be seen, the head node for the verb is unspecified for lemma and word form. It is also unspecified for the grammatical features of the main verb which has to be in present tense. The auxiliary and the main verb together build an analytical word form that is in future tense. On the right side, the following information is given: the result from the composition of the auxiliary catena, the word form catena and the valency frame catena. The resulting verb catena is for the string (bg) *ще бягаш от отговорност* *ste byagash ot otgovornost* (lit. will-FUT run-2SG.PRS (from responsibility-SG.F)) ‘you will run from responsibility’.

Coming back to modeling MWEs and their representation in the lexicon and their realization in the text, we model them in the lexicon as described above assigning an appropriate catena for the forms of the MWE in the paradigm and catena for the valency frame. The realization in the text is performed by the operations defined in the section above. We also represent the grammatical features over two layers: one for the components of the MWE as they appeared in the lexicon, and one for the realization in the text. In the next section we present a classification of the different types of MWEs included in the final integrated lexicon.

6 Analyses of MWE types

In our previous research we gave credit to the most frequent head-based types of MWEs (this means that the MWE is analysed according to its syntactic head – noun, verb, etc.) as presented in BTB-WN. The influence of BTB-WN mapping to the English wordnet also played a big role. When transferred from English, the resulted MWEs in Bulgarian might include free phrases, collocations, etc. to ensure the correct relation to the English notion.

Here we would like to present our model with respect to the complexity of the MWE representation. We view complexity in the following way: a) from fixedness towards flexibility. Here several options are considered: morphological flexibility, syntactic flexibility, semantic flexibility, and combination of two or all of them; b) from continuity to discontinuity. We consider MWEs with at least two words. Please note that the named entities are not discussed. We assume that the more words constitute the MWE, the more complex this MWE is. Idiomaticity is hidden in fixedness. Here are the types we consider: fixed, continuous; fixed, discontinuous; semi-fixed, continuous; semi-fixed, discontinuous; flexible, continuous; flexible, discontinuous.

It can be seen that the fixed, continuous type is mainly nominal or prepositional while the fixed, discontinuous type is rare. The most frequent type is the semi-fixed one. In the continuous subtype noun phrases prevail while in the discontinuous one verbal MWEs are typical. We build on the representation described in Simov & Osenova (2015a,b). Let us consider them in order below. In the graphical representations below we present the main word forms in the paradigm, instead of complete lexical entries.

6.1 Fixed, continuous

Here three main structural variants are detected. They are all idiomatic.

- (1) Noun Conj Noun: (bg) живот и здраве *zhivot i zdrave* (lit. life-SG.M and health-SG.N) ‘some day’ – see Figure 9
- (2) Prep NP:
 - a. (bg) за вечни времена *za vechni vremena* (lit. for eternal-PL times-PL) ‘forever’;
 - b. (bg) между другото *mezhdru drugoto* (lit. between other-SG.DEF) ‘by the way’;
 - c. (bg) на легло *na leglo* (lit. on bed-SG.N) ‘ill’
- (3) Adjective Noun: (bg) добро утро *dobro utro* (lit. good-SG.N morning-SG.N) ‘good morning’

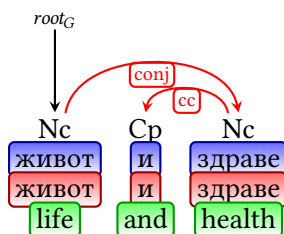


Figure 9: Catena for fixed, continuous expressions: (bg) живот и здраве *zhivot i zdrave* (lit. life-SG.M and health-SG.N) ‘some day’.

The new additions to the catena representation in comparison to our previous work are: the incorporation of the synonyms to the idioms as in examples 1 and 2, and the handling of pragmatic formulae in example 3.

A challenge that appears in this group are the boundaries of the MWEs. For example, (bg) на легло *na leglo* (lit. on bed-SG.N) ‘ill’ might be extended also to the inclusion of a copula: (bg) на легло съм *na leglo sam* (lit. on bed-SG.N am-1SG) ‘to be ill’. The question is whether the copula element should be represented as a component of the MWE or not. According to our suggestion the catena (bg) на легло *na leglo* (lit. on bed-SG.N) ‘ill’ can combine with the catena of the auxiliary and form another catena.

6.2 Fixed, discontinuous

This class is a speaker strategy rather than a distinct type of its own. The strategy can contextualize a fixed MWE and thus add to it more elements. For example, the MWE (bg) без капка разум *bez kapka razum* (lit. without drop-SG.F sense-SG.M) ‘without an iota of sense’ can be extended with a modifier to the noun ‘sense’ such as (bg) без капка медицински разум *bez kapka meditsinski razum* (lit. without drop-SG.F medical-SG.M sense-SG.M) ‘without an iota of medical sense’ in a specific context. These cases are rare and non-systematic.

6.3 Semi-fixed, continuous

This predominantly nominal group contains terms, idiomatic expressions as well as every-day-life expressions. However, its main specificity is the fact that they do exhibit morphosyntactic varieties such as changes in definiteness and number but on the head word only. The dependant remains unchanged.

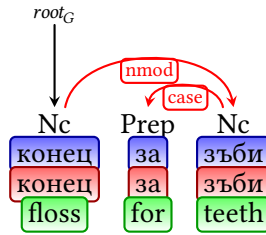


Figure 10: Catena for semi-fixed, continuous expressions: (bg) *конец за зъби* *konets za zabi* (lit. floss-SG.M for teeth-PL) ‘dental floss’.

1. Noun Noun: (bg) *муха цеце* *muha tsetse* (lit. fly-SG.F tsetse) ‘tsetse fly’; (bg) *ангел хранител* *angel hranitel* (lit. angel-SG.M guardian-SG.M) ‘guardian angel’
2. Noun prep Noun: (bg) *конец за зъби* *konets za zabi* (lit. floss-SG.M for teeth-PL) ‘dental floss’ – see Figure 10; (bg) *лак за нокти* *lak za nokt* (lit. polish-SG.M for nails-PL) ‘nail polish’; (bg) *яйце на очи* *yaytse na ochi* (lit. egg-SG.N on eyes-PL) ‘a fried egg’

6.4 Semi-fixed, discontinuous

This group contains mainly verbal MWEs. These are: the quasi-reflexive verbs (the so-called middle verbs where the participating reflexive has no semantics but only a derivational function), and the light verb constructions.

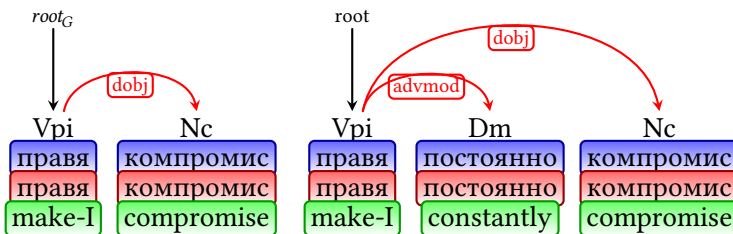


Figure 11: Catena for a light verb construction (semi-fixed, discontinuous expressions): (bg) *права компромис* *pravya kompromis* (lit. do-1SG.PRS compromise-SG.M) ‘to make a compromise’. On the left side is the lexical catena. On the right side is a modification with an adverb, which is realized between the two parts of the MWE.

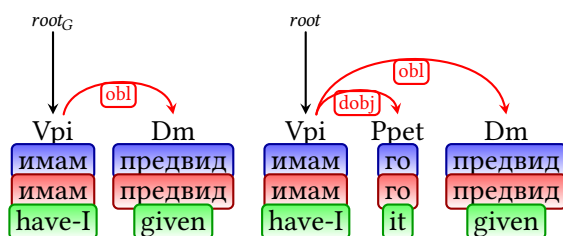


Figure 12: Catena for a light verb construction (semi-fixed, discontinuous expressions): (bg) *имам предвид* *imam predvid* (lit. have-1SG.PRS given) ‘to have in mind’. This is similar to the previous example, but the intervening material is a pronoun.

1. Quasi-reflexive verbs: (bg) *адаптирам се* *adaptiram se* (lit. adapt-1SG.PRS REFL) ‘to adapt’; (bg) *вкисвам се* *vkisvam se* (lit. get-sour-1SG.PRS REFL) ‘to feel bad’
2. Light verb constructions: (bg) *правя компромис* *pravya kompromis* (lit. do-1SG.PRS compromise-SG.M) ‘to make a compromise’ – see Figure 11; (bg) *правя почивка* *pravya pochivka* (lit. do-1SG.PRS rest-SG.F) ‘to take a break’; (bg) *давам обещание* *davam obeshtanie* (lit. give-1SG.PRS promise-SG.N) ‘to make a promise’; (bg) *вкарвам в употреба* *vkarvam v upotreba* (lit. implement-1SG.PRS in usage-SG.F) ‘to put into use’; (bg) *имам предвид* *imam predvid* (lit. have-1SG.PRS given) ‘to have in mind’ – see Figure 12; (bg) *давам под наем* *davam pod naem* (lit. give-1SG.PRS under rent-SG.M) ‘to rent out’

The two parts of the quasi-reflexive verbs can be discontinued by the auxiliary in some forms in the verb paradigm ((bg) *адаптирал съм се* *adaptiral sam se* (lit. adapt-PTCP.PST am-1SG.PRS REFL) ‘I have adapted’). Most of the light verbs have single verbs as synonyms. For example, (bg) *давам обещание* *davam obeshtanie* (lit. give-1SG.PRS promise-SG.N) ‘to make a promise’ has a synonym (bg) *обещавам* *obeshavam* (lit. promise-1SG.PRS) ‘to promise’. They also can often be discontinued by a modifier on the noun element ((bg) *давам голямо обещание* *davam golyamo obeshtanie* (lit. give-1SG.PRS big-SG.N promise-SG.N) ‘to make a big promise’) or by another participant in the sentence (bg) *давам насила обещание* *davam nasila obeshtanie* (lit. give-1SG.PRS reluctantly promise-SG.N) ‘to make a promise reluctantly’). The variant (bg) *давам под наем* *davam pod naem* (lit. give-1SG.PRS under rent-SG.M) ‘to rent out’ allows for an object

coming after the verb (bg) давам *davam* (lit. give-1SG.PRS) ‘to give’: (bg) давам стаята под наем *davam stayata pod naem* (lit. give-1SG.PRS room-SG.F.DEF under rent-SG.M) ‘to rent out the room’.

6.5 Flexible, continuous

This group consists of just one nominal type which is “Adjective Noun”. Some of the MWEs are literal, and some are figurative. In the examples below the last one is figurative.

(4) Adjective Noun

- a. (bg) бежански лагер *bezhanski lager* (lit. refugee-SG.M camp-SG.M) ‘a refugee camp’ – see Figure 13;
- b. (bg) гол охлюв *gol ohlyuv* (lit. naked-SG.M snail-SG.M) ‘a slug’;
- c. (bg) домашна работа *domashna rabota* (lit. home-SG.F work-SG.F) ‘homework’;
- d. (bg) ахилесова пета *ahilesova peta* (lit. Achilles’-SG.F heel-SG.F) ‘Achilles’ heel’

Here the MWEs are mostly terms or near-terms. Both elements form a concept, so they cannot be discontinued but they are flexible with respect to their morphosyntactic behaviour. They can be used with an article or in a plural form. The article occurs only once in a phrase but both elements in the MWE can inflect in number. Also, the idiomatic expressions like (bg) ахилесова пета *ahilesova peta* (lit. Achilles’-SG.F heel-SG.F) ‘Achilles’ heel’ have synonyms, in this case – weakness.

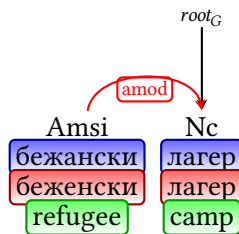


Figure 13: Catena for flexible, continuous expressions: (bg) бежански лагер *bezhanski lager* (lit. refugee-SG.M camp-SG.M) ‘a refugee camp’.

6.6 Flexible, discontinuous

Here some verbal expressions are listed which are flexible with respect to morphosyntax. This means that the verb can inflect in all verb tenses and other verb forms.

(5) Verb NP

- (bg) развързвам кесията *razvarzvam kesiyata* (lit. untie-1SG.PRS purse-SG.F.DET) ‘I pay generously’ – see Figure 14;
- (bg) играя открито *igraya otkrito* (lit. play-1SG.PRS openly) ‘I play fair’;
- (bg) избирам страна *izbiram strana* (lit. choose-1SG.PRS side-SG.F) ‘to take side’;
- (bg) тегля един бой *teglya edin boy* (lit. drag-1SG.PRS one fight-SG.M) ‘to draw a fight’, etc.

The MWE can be used also without the reflexive particle. At the moment we view both possibilities as synonyms. These expressions also allow for some discontinuous material. For example, an adverbial of manner can come between the verb and the object in the first listed MWE above – (bg) развързвам си сериозно кесията *razvarzvam si seriozno kesiyata* (lit. untie-1SG.PRS REFL seriously purse-SG.F.DET) ‘I pay very generously’ – the second tree in Figure 13.

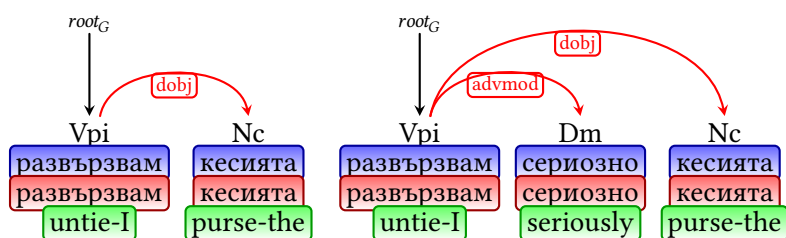


Figure 14: Catena for flexible discontinuous expressions: (bg) развързвам кесията *razvarzvam kesiyata* (lit. untie-1SG.PRS purse-SG.F.DET) ‘I pay generously’.

In this section various examples were outlined according to a proposed classification that respects the complexity of the MWEs. The catena illustrations follow the Universal Dependencies guide.⁹ The fixed, discontinuous type turned out to be a strategy where the speaker can personalize fixedness and thus legitimate the addition of new elements in a specific context.

7 Conclusions and future work

The representation of MWEs in an integrated model has never been a trivial task. Our proposal is to use the catena notion since it allows for a graph-based realization where all the characteristics of interest can be added: the internal structure specifics as well as the external ones, if needed. In addition, the interaction among morphology, syntax (including valency potential and a vanilla mechanism¹⁰ for word order) as well as semantics can be illustrated. We are aware of the fact that our model is similar in many aspects to the other tree-based approaches. At the same time, our representation model is put in the context of an integrated resource and we believe that here come the main novel directions in our work.

It has become clear for quite some time that MWEs are a phenomenon that is not always trivial to define, classify, annotate, analyse and integrate. For that reason, we view our work as a bottom-top effort that would gradually cover specific lemmas, meanings and cases.

Our future work is envisaged in several directions: to fully implement the suggested mechanism, to evaluate it on downstream tasks, and also in the backward direction – to identify the problematic places and repair them in the lexicon. Some already identified problematic places are the MWE boundaries and the degree of granularity in their representation.

Abbreviations

BTB	Bultreebank	IRV	inherently reflexive verbs
BTB-WN	Bultreebank Wordnet	LC	lexicon catena
BVL	Bulgarian Valency Lexicon	LFG	Lexical-Functional Grammar
ID	identifier	MWE	multiword expressions
ILB	Inflectional lexicon of Bulgarian	NLP	Natural Language Processing
		POS	part-of-speech

⁹<https://universaldependencies.org/guidelines.html>

¹⁰This means that our approach is very standard and basic, initially predicting the clear places of discontinuity on the encountered examples without ensuring that all cases are covered appropriately.

SM	semantics	VPC	verb-particle constructions
VMWE	verbal multiword expressions	WFC	Word Form Catena
VFC	Valency Frame Catena		

Acknowledgements

The reported work has been partially supported by CLaDA-BG, *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH*, funded by the Ministry of Education and Science of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure). We would also like to thank the reviewers for their valuable and insightful comments.

References

- Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*. <https://api.semanticscholar.org/CorpusID:13075323>.
- Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 69–108. Language Science Press. DOI: 10.5281/zenodo.2579037.
- Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(4). 23–39. DOI: 10.1007/s10579-009-9094-z.
- Groß, Thomas. 2010. Chains in syntax and morphology. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto & Yasunari Harada (eds.), *Proceedings of the 24th Pacific Asia conference on language, information and computation*, 143–152. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. <https://aclanthology.org/Y10-1018>.
- Laskova, Laska, Petya Osenova, Kiril Simov, Ivajlo Radev & Zara Kancheva. 2019. Modeling MWEs in BTB-WN. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019)*, 70–78. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5109.

- Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.
- Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.
- Lion-Bouton, Adam, Agata Savary & Jean-Yves Antoine. 2023. A MWE lexicon formalism optimised for observational adequacy. In Archana Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han & Shiva Taslimipoor (eds.), *Proceedings of the 19th workshop on multiword expressions (MWE 2023)*, 121–130. Dubrovnik, Croatia: Association for Computational Linguistics. <https://aclanthology.org/2023.mwe-1.16>.
- Masini, Francesca. 2019. *Multi-word expressions and morphology*. Oxford: Oxford University Press.
- O’Grady, William. 1998. The syntax of idioms. *Natural Language and Linguistic Theory* 16. 279–312.
- Osborne, Timothy, Michael Putnam & Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax* 15(4). 354–396. DOI: 10.1111/j.1467-9612.2012.00172.x.
- Osenova, Petya. 2010. Bulgarian. In *The languages of the new EU member states*, vol. 88 (Revue Belge de Phologie et D’Histoire 3), 643–668.
- Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC’14)*, 2785–2792. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/279_Paper.pdf.
- Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Van Der Plas, Behrang Qasemizadeh, Carlos Ramisch, Federico Sangati, Ivelina

- Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471590.
- Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 455–461. Reykjavik, Iceland: European Language Resources Association (ELRA). <https://aclanthology.org/L14-1433/>.
- Simov, Kiril & Petya Osenova. 2014. Formalizing MultiWords as catenae in a treebank and in a lexicon. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova & Adam Przepiórkowski (eds.), *Proceedings of the thirteenth international workshop on Treebanks and Linguistic Theories (TLT13)*, 198–207. Tübingen: University of Tübingen.
- Simov, Kiril & Petya Osenova. 2015a. Catena operations for unified dependency analysis. In Joakim Nivre & Eva Hajičová (eds.), *Proceedings of the third international conference on dependency linguistics (depling 2015)*, 320–329. Uppsala, Sweden: Uppsala University. <https://aclanthology.org/W15-2135>.
- Simov, Kiril & Petya Osenova. 2015b. Modeling lexicon-syntax interaction with catenae. *Journal of Cognitive Science* 16(3). 287–322. DOI: 10.17791/jcs.2015.16.3.287.
- Skoumalová, Hana, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondříčka & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631.
- Vondříčka, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112. 83–101. <https://ufal.mff.cuni.cz/pbml/112/art-vondricka.pdf>.
- Zampieri, Nicolas, Carlos Ramisch & Geraldine Damnati. 2019. The impact of word representations on sequential neural MWE identification. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the joint workshop on multiword expressions and*

Petya Osenova & Kiril Simov

WordNet (MWE-WN 2019), 169–175. Florence, Italy: Association for Computational Linguistics. <https://aclanthology.org/W19-5121>.

Chapter 5

A FrameNet approach to deep semantics for MWEs

👤 Voula Giouli^a, Vera Pilitsidou^b & Hephestion Christopoulos^b

^aInstitute for Language and Speech Processing, ATHENA Research Center, Greece ^bNational and Kapodistrian University of Athens, Greece

We present work aimed at enhancing a semantic lexical resource for Modern Greek with multiword expressions and at manually annotating a corpus with semantic roles with a view to supporting the lexical encoding with corpus evidence. The research was conducted within a larger initiative to construct a Greek FrameNet and corresponding corpus. The ultimate purpose was to provide a shallow semantic representation for multiword lexical units that is similar to the semantic representation of single-word predicates. We focus on both verbal and nominal multiword predicates. Specifically, we address the following questions: (a) what discrepancies seem to be prevalent between single- and multiword entries that are classified under the same frame (in terms of the realisation of Frame Elements), and (b) how to encode these discrepancies.

1 Introduction

Multiword expressions (MWEs) are word combinations that present morphological, lexical, syntactic, semantic, and pragmatic idiosyncrasies (Gross 1982, Baldwin & Kim 2010). In terms of meaning, they do not abide by the semantic interpretation rules of the language by which the meanings of phrases can be constructed out of the meanings of their constituents. In this respect, they appear on a continuum of compositionality: some expressions are analyzable (in that one can “analyze” their constituents in order to understand their meaning), whereas



others are partially analyzable or ultimately non-analyzable at all (Nunberg et al. 1994). The mismatch between their phrasal structure and their deep semantics renders them “a pain in the neck for Natural Language Processing” (Sag et al. 2002). In that regard, the community has been spending considerable effort to model them in a way that facilitates their robust treatment with a view to various applications. However, most MWE-specific lexical resources focus only on the representation of their lexical, morphological, and syntactic properties. Similarly, although several annotated corpora have been developed with the view to training and evaluating algorithms for MWE discovery and classification, little work has been devoted to their semantic representation in corpora with respect to developing applications that require deep semantics. Through our work, we seek to bridge this gap by providing a semantic representation for MWEs in a frame-based lexical resource for Modern Greek.

The chapter is structured as follows: in Section 2 we present the rationale, main objectives, and scope of our work; Section 3 gives an account of the theoretical framework within which our work is placed, as well as previous work on MWEs and their representation in large lexical resources and corpora. Section 4 outlines the methodological principles adopted for creating a frame-based lexical resource for Modern Greek and for treating MWEs. The MWEs that belong to the grammatical categories of noun and verb and their treatment within frames are presented in Section 5. In Section 6, we discuss our findings from the annotation we performed focusing on the discrepancies between single and multiword predicates. Finally, in Section 7, we outline our conclusions and plans for future research.

2 Main objectives

In this chapter, we present work aimed at (i) enhancing a semantic lexical resource for Modern Greek with nominal and verbal MWEs and (ii) manually annotating a corpus with attestations of the lexical units to the end of supporting the lexical encoding with further corpus evidence. The research was conducted within a larger initiative to construct a Greek FrameNet (FN-el) and corresponding corpus (Giouli et al. 2020, Pilitsidou & Giouli 2020). The main objective is to provide a semantic representation for MWEs in a way that is comparable to the one provided for single-word predicates. The goal was to develop a lexical resource coupled with corpus annotation that also treats complex predicates of various kinds; the resource will be useful for numerous Natural Language Processing (NLP) applications. Therefore, to better account for the deep semantics of

complex predicates, we wanted to define their argument structure and provide their lexical-semantic descriptions within the theoretical framework of frame semantics. Our dataset comprises a list of nominal and verbal MWEs extracted from corpora and existing resources in Modern Greek. In the paper, we give an account of their encoding by assigning them to a frame and defining their arguments along with the semantic roles they assume. The construction of the lexicon is based on corpus evidence and the performed annotation.

Finally, in our study, we address two questions: (a) What discrepancies seem to be prevalent between single- and multiword lexical units that are classified under the same frame in terms of Frame Elements assignment and syntactic realization? and (b) How are these discrepancies reflected in the encoding of MWEs and single-word predicates? In other words, what are the discrepancies between, for instance, the single word lexical unit (el) αποφασίζω *apofasizo* ‘to decide’ and the MWE (el) παίρνω απόφαση *perno apofasi* (lit. ‘take decision’) ‘to decide’ in terms of the Frame Elements that are realized? We will demonstrate that the differences between synonymous single- and multiword predicates involve not only variations in the syntactic realization of their (core and non-core) Frame Elements but also in the number of Frame Elements realized. Overall, analyzing these discrepancies might provide insights into how the choice between using a single word predicate and a MWE can influence the syntactic and semantic structure of a sentence, thereby impacting the realization of Frame Elements.

3 Theoretical framework and previous work

Our work draws upon the theory of *Frame Semantics* (Fillmore 1976, 1977, 1982, 1985) as well as the principles and methodologies established by pioneering research in lexical resources, that is inspired by the theory. Frame Semantics is an approach that does not rely on relations like hyperonymy and homonymy, but rather, draws upon the whole of human experience in order to organise the lexicon of any given language. This cognitive approach to the representation of meaning is based on the assumption that, in order to comprehend the meaning of any given utterance, one has to draw on their own experience and knowledge, thus evoking certain schemata. The theory focuses on the continuity that exists between language and experience (Petrucci 1997). In this context, words gain their meaning within a semantic *frame*. A semantic frame schematises an event or a relation, encompassing a system of interconnected meanings. Understanding any one meaning within the frame necessitates grasping all the others. Thus, when any element of this frame is evoked in text or discussion, all other elements become accessible automatically (Fillmore 1982).

Based on Fillmore's theory, the Berkeley FrameNet (BFN, Baker et al. 1998) is a general-purpose lexical semantic resource for English, and it is the earliest and most complete attempt to organise and categorise lexical units in a lexicon based on frames. Frames are seen, thus, as conceptual structures describing specific types of objects, events, or states along with their components, the so-called *Frame Elements* (FEs) of the frame (Baker et al. 1998, Ruppenhofer et al. 2016), whereas the words that evoke a semantic frame are the *Lexical Units* (LUs) of that frame and are unique pairings of a word form and a meaning. Polysemous words typically evoke different frames. LUs pertain to the grammatical categories of verb, noun, adjective, or adverb. In other words, BFN provides a semantic representation that uses frames (or scenes) as its core, and LUs are ultimately organised around frames. Each frame is defined via a gloss that roughly describes the scene represented and a set of FEs; the latter are usually referred to in the gloss. FEs correspond to *semantic roles* specifically defined within each frame and provide finer distinctions of meaning compared to standard semantic roles. The resulting frame annotation scheme is therefore fine-grained. For each frame, the core FEs are generally assumed as central to the meaning conveyed by the frame. Frames are then populated with lexical units (LUs) – both single- and multiword ones. BFN is therefore a means for the semantic representation of LUs within frames regardless of the grammatical category they belong to (noun, adjective, verb, adverb). A set of typed frame-to-frame relations are used to link frames to one another, giving BFN a net-like structure, and – to some extent – a hierarchical organisation. Figure 1 depicts the frame Lending, its FEs – both core (i.e., LENDER, BORROWER, and THEME) and non-core (i.e., DURATION, TIME, PURPOSE, etc) – and the LUs that evoke the frame. A definition of the frame is provided as well as definitions for all the FEs.

Besides English, various FrameNets have been developed for other languages, for example, Japanese (Ohara et al. 2003, Saito et al. 2008), Chinese (You & Liu 2005), German (Erk et al. 2003, Boas 2002), Brazilian Portuguese (Salomão 2009, Timponi Torrent & Ellsworth 2013), Spanish (Subirats 2009), Italian (Lenci et al. 2010), Swedish (Borin et al. 2010), French (Candito et al. 2014), Hebrew (Hayoun & Elhadad 2016), Korean (Kim et al. 2016), Finnish (Lindén et al. 2017), and Modern Greek (Giouli et al. 2020, Pilitsidou & Giouli 2020). In this context, a rather recent initiative, namely, the Global FrameNet Shared Task (Timponi Torrent et al. 2018) seeks to investigate whether frames are universal – and to what extent – and whether BFN can cover the needs of most languages.

Similar to the general-purpose frame-based resources, other domain-specific ones have been implemented depicting language for specific purposes. For example, the language of sports and football has been modeled within the frame

Definition	The Lender gives the Theme to the Borrower with the expectation that the Borrower will return the Theme to the Lender after a Duration of time. This frame differs from the Borrowing frame in that this frame profiles the Lender in active sentences, whereas the Borrowing frame profiles the Borrower.
Example	I lent my girlfriend my car for the weekend
FEs - Core	
Borrower [Borr]	The person or institution who receives the Theme from the Lender for a Duration .
Lender [Lend]	The person or institution who gives the Theme to the Borrower for a Duration .
Theme [Th]	The object that is transferred from the Lender to the Borrower for a Duration .
FEs - Non-Core	
Duration [Dur]	The amount of time in which the Borrower has possession of the Theme .
Manner [Man]	The way in which the Lender lends the Theme .
Semantic Type: Manner	
Place [Pla]	The location in which the Lender lends the Theme to the Borrower .
Semantic Type: Location	
Purpose [Purp]	The aim of the Lender which they believe will be accomplished by lending the Theme to the Borrower .
Time [Time]	The time when the lending event occurs
Frame-frame Relations	
	Inherits from: Giving
	Perspective on: Temporary transfer scenario
Lexical Units	<i>lend.v, loan.n, loan.v</i>

Figure 1: The frame Lending in BFN

semantics paradigm in the so-called Kicktionary database (Schmidt 2009), as well as the Copa-2014 FrameNet Brasil, a frame-based trilingual electronic dictionary covering the domains of Football, Tourism, and the World Cup in three languages, namely, English, Spanish and Brazilian Portuguese (Timponi Torrent et al. 2014); similarly, the BioFrameNet database is a lexical resource built around frames in the domain of molecular biology (Dolbey et al. 2006), whereas frameNets tailored to model the legal (Venturi et al. 2009), financial (Pilitsidou & Giouli 2020) or aviation (Ostroški Anić & Brač 2022) domains have also been developed for languages other than English. Going further, FrameNets that are capable of taking other semiotic modes as input data, for example pictures, and videos have recently been implemented (Timponi Torrent et al. 2022).

The theory of Frame Semantics has been further utilised for the formulation of the Frame-based Terminology (FBT) theory (Faber 2011, 2015) and for the concomitant creation of frame-based terminological databases, like Ecolexicon (Faber & Buendía Castro 2014). Being a cognitive approach to terminology that is based on frame-like representations in the form of conceptual templates underlying the knowledge encoded in specialised texts, FBT directly connects specialised knowledge with Cognitive Linguistics and Semantics (Faber 2015). Specialised language concepts cannot be activated in isolation unless they are part of a larger structure or event. Our knowledge about a concept initially gives us the context or the event in which the concept retains its meaning. In this approach, frames are viewed as situated knowledge structures and are linguistically reflected in the lexical relations that arise from terminographic definitions. Concepts within a thematic field are thus inter-connected with each other based on the events of the field and the frames evoked. These frames are the context in which FBT specifies the semantic, syntactic, and pragmatic behavior of specialised language units. Consequently, instead of being described as static entities out of context, concept representations are treated as dynamic entities within the relevant context (Faber 2011).

Our work builds on the theory of Frame Semantics, Frame-based Terminology, and prior work on BFN, to create a lexical resource that incorporates LUs and frames that belong to language for general purposes (LGP) as well as to language for specific purposes (LSP). To elaborate, we have dealt so far with the grammatical categories of verbs and nouns. Both single and multiword entries have been included in the resource. It is worth mentioning that the majority of the MWE nouns in this work belong to LSP, in other words, they are terms, that is, lexical items characterised by their reference to a scientific field and constitute the (specialised) vocabulary of that field (Sager 1990).

3.1 MWEs in lexical resources

Two types of lexical resources may be identified with respect to MWEs: MWE-dedicated, that is, resources that have been developed with a primary focus on modeling MWEs, and MWE-aware ones that take MWEs into account in addition to other lexical units. Most MWE-dedicated lexical resources are primarily focused on the encoding of their lexical, morphological, and syntactic idiosyncrasies. Recommendations for representing MWEs in mono- and multilingual computational lexica (Calzolari et al. 2002, Copestake et al. 2002) aim at creating a shared model that is suitable for representing MWEs across different languages – yet, they focus mainly on the syntactic and semantic properties of support verbs and noun compounds and their proper encoding thereof. Similarly, Villavicencio et al. (2004) discuss the requirements for the efficient representation of English idioms and verb-particle constructions (VPCs) in lexica by means of augmenting existing single-word dictionaries with specific tables.

In this regard, within the Lexicon-Grammar framework (Gross 1975), French verbal MWEs were classified in the so-called Lexicon-Grammar tables (Gross 1982), where their syntactic and distributional properties and selectional restrictions were represented formally. In this approach, the surface structure of a verbal MWE is represented as a Part-of-Speech sequence of constituents, either continuous or not. The labels N, A, Adv, and PP are used to denote non-lexicalised constituents headed by a Noun, Adjective, Adverb, or Preposition respectively. Lexicalised elements are denoted as *C*. Modification, possible alternations, and distributional properties are encoded as binary properties within the Lexicon-Grammar tables. Along the same lines, similar lexical resources based on the same formal principles and linguistic criteria have been created for verbal idiomatic expressions in other languages, including Greek (Fotopoulou 1993, Mini 2009). The same approach has been adopted for the representation of adverbial MWEs in French by Laporte & Voyatzi (2008) and nominal MWEs in Greek by Anastasiadis-Symeonidis (1986).

Over the years, MWE-specific lexicons of various types have provided elaborate linguistic information for morphological, structural, and lexical properties of MWEs including variation and internal modification of MWEs. Shudo et al. (2011) report on the representation of Japanese MWEs in a comprehensive dictionary that provides detailed descriptions of their syntactic structure (dependencies), internal modification, and functional information. Similarly, Zaninello & Nissim (2010) propose a representation of MWEs in Italian based on their morphosyntactic properties and lexico-semantic information acquired semi-automatically from corpora. Odiijk (2013) reports on the successful experiments and semi-automatic

expansion of DuELME (Grégoire 2010), a lexical database for Dutch MWEs; in the database, MWEs are classified in the so-called equivalence classes based on their syntactic structure, seen as syntactic patterns that occur frequently in a dependency parsed corpus of Dutch.

Recently, MWE-aware lexical resources provide elaborate representations of the structure of MWEs (cf. Leseva et al. 2024, Markantonatou et al. 2024 [this volume]) by making use of the Universal Dependencies formalism (Nivre et al. 2016). Similarly, the notion of the catena provides a mechanism for representing the structure of MWEs (cf. Osenova & Simov 2024 [this volume]). All these representations are aimed at the development of reliable gold standards to aid the task of MWE identification in running text.

In contrast, semantic MWE-aware lexicons, for example, WordNet (Fellbaum 1998), Verbnets (Kipper et al. 2008), SAID (Kuiper et al. 2003), and WikiMwe (Hartmann et al. 2012) give an account of various types of MWEs – yet they are solely focused on their semantic representation, overlooking other aspects. More recently, VerbAtlas (Di Fabio et al. 2019), a large-scale handcrafted lexical-semantic resource aimed at bringing together all verbal synonym sets from WordNet into semantically coherent frames, also treats verb-particle constructions (i.e., *take off*) as well as fully lexicalised idiomatic expressions (i.e., *kick one's heels*, *take a firm stand*, etc.), one of its main contributions being the definition of a set of explicit and cross-frame semantic roles that are linked to the selectional preferences of the verbal predicates.

Moreover, Fotopoulou et al. (2014) propose a model for encoding MWEs of all grammatical categories (noun, verb, adjective, and adverb) providing information on their syntactic structure, morphological and grammatical idiosyncrasies, variation, as well as information about their degree of fixedness. In addition, they provide lexical semantic relations (i.e., synonymy, antonymy, part-hole) giving an account of idiomatic expressions that also bear a literal meaning. To further account for the properties of Greek verbal MWEs, Markantonatou et al. (2019) have developed an infrastructure that accounts for the variability attested and the need for maximal generalisation.

3.2 MWEs in corpora and the corpus-lexicon interface

Besides lexical resources, the modeling of MWEs (i.e., their variations, internal modification, etc.) has also been attempted in both MWE-dedicated and MWE-aware corpora. Notably, the PARSEME initiative features corpora in more than 26 languages from different families that bear annotations for verbal MWEs (VMWEs) facilitating their discovery and identification in running text (Savary

et al. 2017, Ramisch et al. 2018, 2020, Savary et al. 2023). The annotation is performed based on guidelines that are as universal as possible, but which still allow for language-specific categories and tests. The DiMSUM 2016 shared task for joint identification and supersense tagging of nominal and verbal MWEs (Schneider et al. 2016) developed training and test data in English (tweets, service reviews, and TED talk transcriptions). Similarly, a MWE-related dataset in English, Portuguese, and Galician was released within the SemEval-2022 Task 2 (Tayyar Madabushi et al. 2022) on multilingual idiomaticity detection: the task was aimed at identifying whether a sentence contains an idiomatic expression, and at representing potentially idiomatic expressions in context based on semantic text similarity.

Other attempts at MWE semantic annotation in corpora include the annotation of MWEs in the Proposition Bank (PropBank), one of the earliest attempts to develop semantically annotated corpora (Palmer et al. 2005). Support verb constructions and idiomatic expressions in PropBank were later assigned one or more semantic role(s) depending on their meaning (Bonial et al. 2014a,b). Support verb constructions in PropBank were treated in two consecutive annotation iterations: initially, the light verbs were annotated as appropriate by selecting (or creating) the relevant support verb roleset; annotation proper was then performed on the predicative noun. However, one of the main drawbacks of PropBank is that the roleset used is too generic, thus leading to inconsistencies in labelling.

In between the corpus and the lexicon, Giouli (2023) proposes a model for representing the semantics of VMWEs by (a) taking into account their inherent idiosyncrasies: lexical, syntactic, and semantic, and (b) linking lexicon entries with their occurrences in a corpus that bears rich linguistic annotations (including Semantic Role Labelling). The model is claimed to entail a holistic approach to VMWE representation.

By default, BFN is placed in the lexicon-corpus and syntax-semantics interface. Therefore, it accounts for the semantics of lexical entries also considering context within frames. This holds true for single and multiword entries. Lexicalised noun-noun compounds (i.e., *wheel chair.n*), verb-particle constructions (i.e., *help out.v*), as well as idiomatic expressions (i.e., *aid and abet.v*, and *cook someone's goose.v*) are treated on their own as LUs, that pertain to the grammatical categories of noun or verb. For example, the verbal MWEs *aid and abet.v* and *help out.v* are both assigned to the frame *Assistance*, and their FEs along with their syntactic realisation are attested as shown in Table 1.

While BFN includes MWEs in the database, it does not analyze them internally. However, sentences in BFN bear a multi-layer annotation: Frame Element, Gram-

Table 1: Encoding of the MWE LU help out.v in BFN

Frame Element	syntactic realisation	n. of occurrences
BENEFITED-PARTY	NP.Obj	3
FOCAL-ENTITY	PP(of).Dep	1
GOAL	DNI	2
HELPER	NP.subj	3

mathematical Function, and Phrase Type, and thus constitute clear examples of basic combinatorial possibilities (valence patterns) for each target LU. In this regard, all BFN annotations are constellations of triples that make up the FE realisation for each annotated sentence, each consisting of a FE or semantic role that is relevant to the frame itself (i.e., Agent, Experiencer, Cogniser, etc.), a grammatical function (i.e., Subject, Object) and a phrase type (i.e., Noun Phrase (NP), Verb Phrase (VP), Prepositional Phrase (PP), etc.). As a result, the syntactic realisation of the FEs is revealed via the annotation performed on the LUs and their FEs. This annotation provides us with a description of the syntactic valence properties of LUs, that is, the syntagmatic types that co-occur in the syntactic locality of the lexical item plus the grammatical functions they assume, as shown in (1):

- (1) [All these commissions_{HELPER}] *helped* [me_{BENEFITED-PARTY}] *out* [of the pains_{FOCAL-ENTITY}]
 [All these commissions.NP-SUBJ] *helped* [me.NP-OBJ] *out* [of the pains.PP]

Building on the dichotomy between the syntactic and semantic heads of expressions, only relatively recently has BFN given an account of the representation of support verb constructions in the database (Petrucek & Ellsworth 2016). In this approach, the semantically empty support verb is assigned the tag *Supp*, whereas both frame assignment and annotation are performed with the predicative noun as the target as shown in (2).

- (2) [Horatio_{PROTAGONIST}] *took*^{Supp} a *dirty nap*. (Petrucek & Ellsworth 2016)

FrameNets for other languages, for example, German, also treat MWEs of various types including support or light verb constructions, idioms, and metaphors (Burchardt et al. 2009). Finally, Borin (2021) discusses the inclusion of MWEs in the Swedish FrameNet++, also elaborating on the description of MWEs from a broad typological point of view. In this study, we elaborate on the idiosyncrasies of MWEs and the issues raised during annotation.

4 Methodology

In this section we present the methodology we adopted for building our frame-based lexical resource, outlining the different steps taken in the development process. It should be noted that the approach taken to FrameNet development is not uniform: teams have adopted various methodologies, ranging from manual construction entirely from scratch (in a way that is similar to the lexicographic process followed in BFN) to projecting translations from BFN to the target language, and even to semi-automatically grouping LUs for creating frames using data-driven techniques. In all these cases, the question raised is whether the frames defined in BFN for the English language are generally applicable to other languages as well, given the cultural differences entailed, as well as the idiosyncrasies and grammatical peculiarities of each language, and how and to what extent mappings from one FrameNet to another are feasible. From another perspective, there are three approaches to frame development (Ruppenhofer et al. 2016, Candito et al. 2014, Virk et al. 2021), namely, the lexicographic frame-to-frame strategy, the corpus-based lemma-to-lemma approach, and the full-text strategy. The lexicographic frame-to-frame strategy is aimed at documenting the range of syntactic and semantic combinatorial possibilities of words in each of their senses. Thus, annotation is performed on selected sentences of the corpus, that is, sentences that best record the valences of words. In this approach, annotation is relative to one lexical unit per sentence: the target. In general, we select sentences for annotation where, with the exception of subjects, all frame elements are realised locally by constituents that are part of the maximal phrase headed by the target word. The frame-by-frame strategy enforces coherence of annotations within a frame (Candito et al. 2014). By contrast, in the full-text annotation mode, all content words, that is, words bearing a lexical meaning, are treated as targets, and annotation is directed toward their dependents. In between the two strategies, the lemma-by-lemma annotation mode is focused on lemmas – possibly polysemous ones – rather than frames, and the annotation of these lemmas within different frames.

Although BFN was constructed as a general framework for applying semantic annotations on textual data cross-linguistically, certain frames need to be adapted to fit other languages. To this end, prior to annotation proper, a pilot annotation phase was carried out (Giouli et al. 2020) in which translations from BFN were projected to the Greek data. As shown in Table 2, in most cases, the BFN frames were applicable to the Greek data. However, we could not account for 12.3% of LUs, due to either a *frame shift* (i.e., a frame change) or a missing frame (i.e., a frame that is not provided for English). Researchers working on other languages also report frame shifts (Yong et al. 2022). To avoid shortcom-

ings and gaps, we opted for constructing the Greek FrameNet manually from scratch instead of projecting annotations.

Table 2: From BFN to FN-el: appropriateness of BFN to Greek.

	number	percent
perfect fit	549	87.70%
non perfect fit	54	8.63%
missing frame	23	3.67%
total	626	100.00%

After a closer inspection of the data, the following reasons for frame shifts were identified (in order of occurrence):¹

- *Too specific*: the LU requires a frame more generic than the one available in the original database;
- *Too generic*: the LU requires a frame more specific than the one available in the original database;
- *Different causative alternation*: the LU requires a causative interpretation that is not present in the original frame, which may be either inchoative or stative;
- *Different inchoative alternation*: the LU requires an inchoative interpretation that is missing in the original frame, which may be either causative or stative;
- *Missing FE*: the original frame lacks a FE that is required in the target frame;
- *Extra FE*: there is a FE in the original frame that is not required in the target frame;
- *Different perspective*: the LU was proved to impose a perspective that is different from the one in the original frame;
- *Different stative alternation*: the LU requires a stative interpretation that is not present in the original frame, which may be either causative or inchoative;

¹These tags were to a great extent adopted from Global FrameNet annotation.

- *Different entailment*: the LU has different entailments from the ones foreseen by the original frame;
- *Different coreness status*: some non-core FE should be core in the target language.

Within the FN-el project, we adopted a modular approach to lexicon development, in the sense that predicates pertaining to a pre-defined set of semantic classes (namely, emotion, cognition, communication) or domains (finance, health) were selected and accounted for, thus opting for a domain-by-domain strategy.² More precisely, micro-projects were run towards treating predicates that pertain to each semantic class and/or domain. In this regard, we adopted the lemma-to-lemma strategy followed by a frame-to-frame one; multiple iterations of this procedure were conducted.

The task was organised as a four-stage procedure: (a) corpus creation and LU selection; (b) frame schematisation based on the syntactic and semantic properties of the selected LUs; (c) corpus annotation with a view to confirming or rejecting our initial intuitive decisions; and (d) frame validation and adjudication, where appropriate, and their extension with new LUs. More precisely, custom-made corpora of newswire texts, as well as corpora with a high term ratio that pertain to specialised domains were created to identify and extract words pertaining to the grammatical categories of noun and verb – also coupled with statistical information. An effort was made to extract the MWEs (verbal and nominal) from the corpora. N-grams were then extracted using SketchEngine (Kilgarriff et al. 2014), whereas terms were extracted semi-automatically using AntConc (Anthony 2005).³

After sense discrimination for polysemous words, meaningful groupings of word-sense pairings were performed – initially based solely on dictionary definitions. Frames were then constructed and populated with LUs; polysemous words fall under different frames, depending on their meaning within a given context. Each frame was further enhanced via the definition of the schema evoked and schematised via its FEs (core and non-core). Stipulating FEs was perhaps the most challenging aspect of the work. Note that core FEs grant a frame its uniqueness. Moreover, relations between frames were defined, the most important being *Inheritance*, *Perspective-on*, *Using*, *Subframe*, and *Precedes*.

²This is the approach taken to the French FrameNet construction (Candito et al. 2014) and is assumed to enforce the coherence of frame delimitations.

³Available online: <http://www.laurenceanthony.net/>.

This procedure for lexicon building is seen as the bottom-up part of the hybrid methodology we adopted: from corpora and lexical units to the definition of frames. The bottom-up approach to lexicon creation process was then complemented with a top-down one, according to which the frames were then populated with new LUs, that is, single- and multiword entries that are synonymous to the existing ones. The two approaches are complementary and were initiated in cycles during the project.

5 The treatment of MWEs in FN-el

Currently, our FN-el database contains c. 2,500 LUs organised around 62 frames. Of these, a total of 561 LUs are terms in the domain of finance, their termhood being determined based on specific criteria; we ended up with 39 frames (9 scenes) for the domain of finance. The remaining LUs are treated under frames in the semantic classes of activity, cognition, communication, and emotion. Numerical data regarding the current status of FN-el is depicted in Table 3.

Table 3: LUs in FN-el: numerical data.

	single	multiword	total
nouns	823	205	1028
verbs	671	572	1243
adjectives	127	32	159
adverbs	84	3	87
total	1705	812	2517

Each frame contains a definition of the scenario (gloss), the FEs (both core and non-core) along with the LUs that populate it. LUs that pertain to the grammatical categories of noun and verb have been extensively treated so far; both single and multiword lexical units are included in the resource and encoded as appropriate. An example of a frame in FN-el, namely, the Agreement - or - Disagreement one is presented in Figure 2. As shown, the gloss (definition) showcases the FEs of the frame – both core and non-core ones; FEs are also coupled with glosses. The LUs that evoke the frame are also provided. In our resource, we retain the respective terminology: names of frames, FEs, frame-to-frame relations, and glosses are all in English. In effect, using English as metadata ultimately facilitates the alignment of FN-el to BFN.

Definition	COGNISER-1 and COGNISER-2 hold a positive or negative opinion with respect to a TOPIC . COGNISER-1 and COGNISER-2 may also be referred to in the text collectively as COGNISERS . COGNISER-1 and COGNISER-2 may also appear as OPINION HOLDER-1 and OPINION HOLDER-2 . The INTENSITY of the opinion expressed, the MANNER of the expression and the REASON may also be expressed.
FEs - Core	
COGNISER	The COGNISER holds an opinion about a particular TOPIC ; this opinion is seen in comparison to the opinion of another COGNISER .
TOPIC	A phenomenon or state or affairs that the COGNISER is considering with respect to their opinion.
OPINION HOLDER	The OPINION HOLDER holds a particular opinion, or point of view, which may be portrayed as being about a particular TOPIC .
FEs - Non-Core	
INTENSITY	Any description of the degree to which COGNISER-1 and COGNISER-2 hold an opinion or point of view about a particular TOPIC .
MANNER [Man] Semantic Type: Manner	Any description of how COGNISER-1 and COGNISER-2 hold and express the same or different opinion about a particular TOPIC .
REASON	Typically, the rationale or motivation behind the opinion held by the COGNISERS . In can be realised as a PP-ως constituent.
Lexical Units	<i>διαφωνία.n diafonia 'disagreement', διαφωνώ.v diafono 'disagree', δίνω τα χέρια.vid dino ta cheria (lit. 'give the hands) 'to agree', κάνω συμφωνία.lvc kano simfonia (lit. 'make agreement') 'to agree' συμφωνία.n simfonia 'agreement', σύμφωνος.adj simfonos 'congruent', συμφωνώ.v simfono 'to agree'</i>

Figure 2: The Agreement-or-disagreement frame in FN-el

MWEs that are listed as LUs in a frame appear in their *canonical form*: for nominal MWEs (NMWEs), that is, MWEs headed by a noun, the canonical form entails that the head noun is in the nominative case, singular number. A VMWE in its canonical form is a verbal phrase whose head verb is in a lemma form and whose other lexicalised components depend either on the verb or on another lexicalised component; non-lexicalised elements and open slots are not included in the canonical form. Since lexicon building is based on pre-processed data, we are no longer interested in the representation of the internal structure of the MWEs and their syntactic variations; these are depicted via the annotated instances that are included as examples in the database. We will elaborate on the treatment of MWEs and the representation of their valences in Sections 5.1 and 5.2.

5.1 Nominal MWEs

So far, 205 NMWEs have been included as LUs in the database and were assigned a frame based on their meaning. Currently, a large portion of the NMWEs encoded in FN-el are terms pertaining to the specialised language of finance and banking (133 LUs out of 205). The NMWEs for the financial domain were extracted semi-automatically from domain corpora using the methodology presented in Section 4. However, since these LUs belong to LSP, we had to diverge from BFN's frames in many ways described below. In terms of their structure, the NMWEs included in FN-el are constructions that have been extensively discussed in the literature on Modern Greek, namely, Adjective Noun (A N), Adjective Adjective Noun (A A N), Noun Noun (N N), and Noun Noun in the genitive (N N_{GEN}) sequences (Anastasiadis-Symeonidis 1986, Ralli 2007, Gavriilidou 2013). In this regard, the NMWE in (3) is an A N construction headed by the N, whereas, the NMWE in (4) falls in the category of N N_{GEN} constructions, where the second, non-head constituent is assigned the genitive case. The NMWE in (5) is an A A N continuous structure, where the third constituent, the noun, functions as the head, while (6) is an example of a N N structure, with its first constituent being the head.

- (3) κόκκινο δάνειο
kokkino danio
red loan
'non-performing loan'
- (4) φόρος εισοδήματος
foros isodimatatos
tax income.GEN
'income tax'

- (5) καθαρά έντοκα έσοδα
 kathara entoka esoda
 net.PL interest.bearing.PL earnings.PL
 ‘net interest income’
- (6) δείκτης DAX
 diktis DAX
 index DAX
 ‘DAX index’

These [A N] and [N NGEN] sequences are LUs with a non-compositional meaning, in that their meaning is not the product of the meaning of their parts. In this regard, the NMWE depicted in (3) is not a loan colored red, but a non-performing one. They are phrasal, and thus syntactic entities, sharing some features with (morphological) compounds, and are inaccessible for the syntactic operations that phrases normally allow. In that respect, they are continuous structures, in the sense that the order of their constituents is fixed, and no other elements can be inserted in between; in some cases, they do not even allow modification. Therefore, as in other lexicographic projects, one of the most challenging issues while creating the resource has been the recognition of NMWEs based on linguistic criteria, and their inclusion in a frame thereof.

Once they were assigned to a frame, the annotation of running text was performed. We aimed to find the syntactic structures MWEs occur in and the valences of MWEs. We will elaborate on the annotation and the issues raised in Section 6. The output of this annotation reveals the FEs that are specific to the LU at hand in the specific frame as well as their syntactic realisations. An example of the representation of a NMWE is provided in Table 4. Namely the multi-word LU (el) κόκκινο δάνειο.nmwe *kokino danio* (lit. ‘red loan’) ‘non-performing loan’ evokes the LENDING frame to which it has been assigned as a LU of the grammatical category nmwe. Its definition (gloss) is provided in Greek as a paraphrase: (el) μη εξυπηρετούμενο δάνειο *mi exipiretumeno danio* ‘non-performing loan’; it has also been assigned FEs as appropriate along with their realisations attested in the annotated corpus.

As shown in (7), the FE BORROWER is realised either as a NP in the genitive or as a PP headed by the preposition σε *se* ‘to’ as shown in (7) and (8) respectively. Once the BORROWER is realised as a NP in the genitive, the FE LENDER is instantiated by a PP headed by the preposition από *apo* ‘by’ as shown in (7); otherwise, it is realised as a NP in the genitive (8):

Table 4: he LU *κόκκινο δάνειο.nmwe* ('non-performing loan') in FN-el.

Frame element	Syntactic realisation	Occurences
BORROWER	NP.Dep	3
BORROWER	PP(σε).Dep	1
LENDER	NP.Dep	1
LENDER	PP(από)	1
AMOUNT	NP.Dep	1
DURATION	PP(για)	1
DURATION	NP.Dep	1
TIME	NP(μέχρι)	2
CAUSE	AJP.Dep	1
CAUSE	N.Dep	1

- (7) *κόκκινα δάνεια* [επιχειρήσεωv_{BORROWER}] [από την ETE_{LENDER}]
 kokkina dania epichiriseon apo tin ETE
 red.PL loan.PL enterprise.PL.GEN from the.SG.ACC NBG.SG.ACC
 'non-performing loans to households from NBG'
- (8) *κόκκινα δάνεια* [τραπεζών_{LENDER}] [σε επιχειρήσειs_{BORROWER}]
 kokkina dania trapezon se epichirisis
 red.PL loan.PL bank.PL.GEN to enterprise.PL.ACC
 'non-performing loans to enterprises from NBG'

Notably, shifts or subtle differences in meaning or differences in perspective between LUs are made evident via their FEs. For example, both the multiword term (el) πιστωτικό γεγονός.nmwe *pistotiko gegonos* (lit. 'credit event') 'bankruptcy' and its near synonym (el) πτώχευση.n *ptochefsi* 'bankruptcy' evoke the frame *Wealth* with *INSTITUTION* and *PERSON* being defined as core FEs of the frame. However, differences in the realisation of FEs shed light on the nuances of the two near-synonymous LUs; as shown in (9) and (10), the LU (el) πτώχευση accepts both *PERSON* and *INSTITUTION* as FEs, whereas the multiword term (el) πιστωτικό γεγονός accepts only *INSTITUTION* as displayed in (11) and (12).

- (9) η πτώχευση [της Thomas Cook_{INSTITUTION}]
 i ptochefsi tis Thomas Cook
 the bankruptcy the.SG.GEN Thomas Cook
 'the bankruptcy of Thomas Cook'

- (10) η πτώχευση [ενός εκ των συζύγων_{PERSON}]
 i ptochefsi enos ek ton sizigon
 the bankruptcy one.SG.GEN of the.PL.GEN spouse.PL.GEN
 ‘the bankruptcy of one of the spouses’
- (11) πιστωτικό γεγονός [για την Ελλάδα_{INSTITUTION}]
 pistotiko gegonos gia tin Elada
 credit event for the.SG.GEN Greece.SG.GEN
 ‘A credit event for Greece’
- (12) * πιστωτικό γεγονός [για τον σύζυγο_{PERSON}]
 pistotiko gegonos gia ton sizigo
 credit event for the.SG.ACC spouse.SG.ACC
 ‘A credit event for the spouse’

5.2 Encoding Verbal MWEs

Following the typology and criteria defined in the PARSEME initiative (Savary et al. 2017, Ramisch et al. 2018, 2020, Savary et al. 2023), four types of verbal MWEs have been included in the resource: (a) verbal idiomatic expressions (VIDs), that bear a meaning that cannot be computed based on the meaning of their constituents and the rules used to combine them, for example, (el) βάζω πλώρη *vazo plori* (lit. ‘put.PRS.1SG prow.SG.ACC’) ‘to set forth’; (b) light verb constructions (LVCs), i.e., expressions with a rather transparent meaning that comprise a support or light verb that is semantically empty and a predicative noun or a predicative adjective or a prepositional phrase, for example, (el) δίνω υπόσχεση *dino yposchesi* (lit. ‘give.PRS.1SG promise.SG.ACC’) ‘to promise’; (c) multi-verb constructions (MVCs), that is, expressions with coordinated lexicalised head verbs, for example, (el) απορώ και εξίσταμαι *aporo ke existame* (lit. ‘wonder.PRS.1SG and be.very.surprised.PRS.1SG’) ‘to be very surprised’; and (d) verb-particle constructions (VPCs) comprising a verb and one of the adverbs (el) μπροστά *brosta* ‘in front, forward’, μπρος *bros* ‘in front, forward’, πίσω *πισο* ‘back’, πάνω *pano* ‘up’, κάτω *kato* ‘down’, μέσα *mesa* ‘in’, έξω *exo* ‘out, outside’ in Greek. These adverbs are not morphologically derived from adjectives and exhibit most, if not all, of the properties that particles in other languages have (Giouli et al. 2019). Moreover, they have two distinct functions: as adverbs denoting time or location, they are used as modifiers; combined with prepositions, they form complex prepositions (Holton et al. 1997), as for example (el) μπροστά από *brosta apo* (lit. ‘in-front from’) ‘in front of’, (el) μέσα σε *mesa se* (lit. ‘in to’) ‘in’, (el) πάνω από *pano apo* (lit. ‘over of’) ‘over’, etc. Given their resemblance with VPCs in other languages

in terms of their properties, we decided to retain the latter class for Greek as well, and therefore expressions as the ones depicted in (13) and (14) were classified as VPCs. In terms of their semantics, VPCs were identified as non-compositional in meaning. As previously shown (Savary et al. 2019), these constructions are the most ambiguous. Depending on the context, they can be used literally and have a fully compositional meaning. In that case, they are not VMWEs.

- (13) *πέφτω*_i *μέσα* *στις* *προβλέψεις* *μου*_i
 pefto mesa stis provlepsis mu
 fall.PRS.1SG in to-the.PL.ACC prediction.PL.ACC my.1SG
 ‘to succeed in my predictions’
- (14) *βάζω* *μπρος* *τη* *μηχανή*
 vazo bros ti michani
 put.PRS.1SG forward the.SG.ACC engine.SG.ACC
 ‘to start the engine’

Once they were selected for inclusion, they were assigned a frame based on their semantics. As mentioned above, we have so far treated VMWEs that belong to the semantic domains of emotion, cognition, and communication – and the respective frames. For example, the LVCs (el) *κάνω μάθημα*.lvc *kano mathima* (lit. ‘make.PRS.1SG lesson.SG.ACC’) ‘to teach’, (el) *δίνω μάθημα*.lvc *dino mathima* (lit. ‘give.PRS.1SG lesson.SG.ACC’) ‘to teach’, (el) *δίνω συμβουλή*.lvc *dino symvuli* (lit. ‘give.PRS.1SG advice.SG.ACC’) ‘to advice’ and (el) *δίνω οδηγία*.lvc *dino odigia* (lit. ‘give.PRS.1SG instruction.SG.ACC’) ‘to instruct’, have been included in the resource within the Transferring-knowledge frame which also includes the single word LUs *διδάσκω*.v *didasko* (‘to teach’), *μαθαίνω*.v *matheno* (‘to teach’), etc. Variants of the selected VMWEs were included in the database as separate LUs and encoded as appropriate. For example, the LVC (el) *παίρνω απόφαση* *perno apofasi* (lit. ‘take.PRS.1SG decision.SG.ACC’) ‘to decide’ and its variant form (el) *λαμβάνω απόφαση* *lamvano apofasi* (lit. ‘take.PRS.1SG decision.SG.ACC’) ‘to decide’ are both treated as LUs in the Deciding frame; the latter has a formal register.

At the next stage, the arguments of the semantic predicate, that is, the VMWE taken as a whole, were identified and assigned FEs as appropriate. In this respect, we are no longer interested in the internal structure of the VMWE, that is, its fixed or lexicalised elements and the grammatical functions they assume, but rather in the non-fixed ones. Thus, FEs realised as arguments or adjuncts of the VMWE (taken as a whole) were identified and encoded.

6 Corpus annotation

Corpus annotation in BFN and related projects is aimed at documenting the range of syntactic and semantic combinatorial possibilities, or valences, of words in each of their senses. FrameNet annotation is always done relative to one particular lexical unit, the target, which is most often a single-word but can also be a multiword expression such as a phrasal verb (for example, *give in*) or an idiom (e.g., *take into account*). In this respect, the final step in our work was the annotation of selected instances of the MWEs used in context. One consideration, therefore, has been the selection of sentences from the corpus that will serve as ideal examples to annotate. This procedure resulted in the validation of frame definition and assignment and led to revisions and amendments where needed. The annotated corpus currently amounts to ca. 2600 sentences.

Annotation was performed on top of textual data that were pre-processed automatically via UDPipe (Straka & Straková 2017) at the levels of lemmatisation, part-of-speech (POS) tagging, and dependency parsing. Annotation on the lexical level was performed manually. Two students annotated selected sentences using the web annotation tool WebAnno (Yimam et al. 2013). Annotation was performed as a two-step procedure taking both verb and noun as targets. At the first stage, MWEs that constitute semantic predicates mapped onto a concept were selected. The selected markables were then annotated at the SemPred layer which is available as a WebAnno built-in module. According to the guidelines set, the markable was assigned a Part-of-Speech tag as appropriate, and the canonical form of the MWE at hand. A second span layer, namely, SemArg, represents slot fillers. The arguments and modifiers of the MWE (taken as a whole) were identified, and the semantic roles they assume were further specified. An instance of the annotation tool is illustrated in Figure 3.

The screenshot shows the WebAnno interface with two sentences. The first sentence is partially visible: "Το τέλος του «Μινιόν» ήρθ... της ελληνικής εκβιομηχάν...". The second sentence is "Ακόμα και αν ο ιδιοκτήτης του έκανε φιλότιμες προσπάθειες, μαζί με τους εργαζομένους να το ξαναστήσει, δεν κατάφερε ποτέ να ξαναγίνει αυτό που ήταν πριν." Annotations include "ο ιδιοκτήτης του" as "Arg0 | Agent", "έκανε" as "vsup", "Μanner", "κάνω προσπάθεια" as "lvc", and "μαζί με τους εργαζομένους να το ξαναστήσει" as "Arg1 | Goal". A tooltip for "ο ιδιοκτήτης του" shows "SemArg" and "ID: 212321". The right sidebar shows the "SemPred" layer with "Text" set to "προσπάθειες", "LU-id" set to "κάνω προσπάθεια", "LU-pos" set to "lvc", and "frame" set to "trying".

Figure 3: Annotating MWEs in Webanno.

Annotations were carried out independently by the two annotators; however, in order to ensure the highest quality of the dataset created, extended discussions followed each annotation cycle and adjudication of the annotations was performed where needed.

At this point, in order to better account for the properties of MWEs in Greek and their idiosyncrasies, a short description of the Greek language is in order. Modern Greek is a highly inflected language: nouns, adjectives, and certain pronouns show a rich inflectional system that features three grammatical genders (masculine, feminine, neuter), singular and plural numbers, and four cases (nominal, genitive, accusative, and vocative). The verbal inflectional system is equally rich: verbs inflect for person, number, tense, aspect, etc. Moreover, in terms of syntax, Greek is a language with a relatively free order of main constituents in a clause. The basic or unmarked order mainly follows the verb-subject-(object) pattern (Holton et al. 1997: 426); however, other variations are also attested, but these alternatives are appropriate in certain discourse contexts (Holton et al. 1997). This flexibility is due to case marking that signals the function of nominals: subjects are attested in the nominative case, whereas objects are most often in the accusative or in the genitive case; nominal complements of prepositions are also either in the accusative or the genitive. Finally, being a pro-drop language, Greek allows null subjects; the absence of a full or weak subject pronoun is accommodated by verbal morphology.

Following the above, MWEs often occur in various configurations. As a guideline, we tried to select sentences for annotation in which all FEs of the frame are realised by constituents that are part of the maximal phrase headed by the target word, including subjects – if possible. It should be noted that the BFN uses the Constructional Null Instantiation (CNI) tag as a mechanism to model the omitted constituents. Cases of CNI include the omitted subject of imperative sentences, the omitted agent of passive sentences, and of course null subjects, or the PRO-elements; we only adopted the afore-mentioned approach for the null subjects in cases where we had to include such sentences in the corpus.

6.1 Annotation with NMWEs as targets

Annotation with NMWEs as targets was relatively easy, as most NMWEs are continuous structures; modifiers of these NMWEs are realisations of their FEs. For example, the NMWEs (el) φόρος εισοδήματος.nmwe *foros isodimatos* (lit. ‘tax income.SG.GEN’) ‘income tax’ and (el) τέλη κυκλοφορίας.nmwe *teli kykloforias* (lit. ‘tax.PL circulation.SG.GEN’) ‘road tax’ which are subsumed under the Tax-payment frame, are annotated as taking the FEs TAXPAYER and AMOUNT, as shown in (15):

- (15) φόρος εισοδήματος [φυσικών προσώπων_{TAXPAYER}]
 foros isodimatos fysikon prosopon
 tax income.SG.GEN natural.PL.GEN person.PL.GEN
 [3,7 δις. ευρώ_{AMOUNT}]
 3.7 disekatomiria evro
 3.7 billion.PL.ACC euro.PL.ACC
 ‘personal income tax amounting to 3.7 billion euros’

In some cases, NMWEs come in the form of structures with shared heads as nested expressions, raising issues during annotation. As they are encoded as separate LUs in the database, annotation uses the feature *Null* retained for non-lexicalised constituents, and annotation is performed for each MWE separately.

- (16) Τα [κόκκινα_{TYPE}] στεγαστικά δάνεια
 ta kokina stegastika dania
 the.PL red.PL home.PL loan.PL
 ‘the non-performing home loans’

When annotation was performed with a verb as targets, occurrences of NMWEs were annotated as FEs of the respective frames. As shown in (17), the NMWE (el) κεντρική τράπεζα.nmwe *kentriki trapeza* ‘central bank’ is realised in the sentence as the BORROWER of the frame Lending in which the LU δανείζω.v *danizo* ‘lend’ occurs, whereas, the NMWE LU (el) εμπορικές τράπεζες *eborikes trapezes* ‘commercial banks’ (headed by the preposition από *apo* ‘by’) is realised as the LENDER.

- (17) [Η κεντρική τράπεζα_{BORROWER}] δανείζεται
 I kentriki trapeza danizete
 The.SG.NOM central.SG.NOM bank.SG.NOM borrow.PRS.3SG
 [χρήματα_{THEME}] [από τις εμπορικές τράπεζες_{LENDER}]
 chrimata apo tis eborikes trapezes
 money.PL.ACC from the.PL.ACC commercial.PL.ACC bank.PL.ACC
 ‘The central bank borrows money from the commercial banks’

6.2 Annotation with VMWEs as targets

Annotation of VMWEs proved to be challenging. Only VMWEs in an idiomatic use were taken into account, whereas literal occurrences of MWEs were not annotated. Literal occurrences of MWEs, also referred to as their literal readings

or literal meanings, have received considerable attention equally from the linguistic and the computational communities. In an experiment run for German, Greek, Basque, Polish, and Brazilian Portuguese, Savary et al. (2019) report almost 11.5% of the VMWE occurrences in the Greek corpus to be literal readings of the VMWE surface forms – a phenomenon referred to as the *literal-idiomatic ambiguity*.⁴ Other VMWEs were found to be semantically ambiguous (17% of the VMWEs), bearing different meanings based on the context. Usually, VIDs that comprise a verb predicate and the weak form of a personal pronoun are ambiguous, whereas LVCs and VPCs were also found to have more than one sense or usage.

In our database, 31 out of the 671 LUs that are VMWEs (4.77%) are also instances of polysemous entries. Following standard lexicographic practices, the latter were subsumed under different frames based on their meaning. For example, the LVC (el) δίνω απάντηση *dino apantisi* (lit. ‘give answer’) ‘to answer’ in (18) has been included in the Communicating-response frame; in a broader sense depicted in (19), it also evokes the Expressing-opinion one. The two frames are defined via two distinct sets of FEs as shown in Table 5.

Table 5: The Communicating a response and Communicating an opinion frames.

Frame	Definition	FEs
Communicating a response	A Speaker uses language (oral or written) to answer a certain Question that might be asked by an Enquirer. The Manner and Means might be mentioned.	Speaker Enquirer Topic Manner
Communicating an opinion	A Speaker or Statement uses language in order to share or make public their Opinion about a certain Topic. Their Strength of Opinion might be present as an adverb.	Speaker Opinion Topic Strength

Once their sense was disambiguated, encoding and annotating them posed no serious problems. Like single-word verb predicates, issues that arise during the annotation of VMWEs of all types are relevant to the granularity of the role-set employed or the specification of the appropriate role. Our approach to MWEs in FN-el is comparable to the approach taken in BFN – especially for the LVCs.

⁴For a definition of the literal-idiomatic ambiguity, see (Savary et al. 2019).

Annotation was performed with the semantic head, that is, the predicative noun, as the target as shown in (18) and (19).

- (18) Η υπουργός έδωσε^{Supp} σαφή απάντηση
 i yurgos edose safi arantisi
 The.SG.NOM minister.SG.NOM give.PST.3SG clear.SG.ACC answer.SG.ACC
 στους μαθητές.
 stus mathites
 to.the.PL.ACC students.PL.ACC
 ‘The minister gave clear answers to the students.’
- (19) Το κείμενο δίνει^{Supp} πειστικές απαντήσεις
 To kimeno dini pistikes arantis
 The.SG.NOM text.SG.NOM give.PRS.3SG convincing.PL.ACC answer.PL.ACC
 σε αιώνια προβλήματα.
 se eonia proulimata
 to eternal.PL.ACC problem.PL.ACC
 ‘The text provides answers to eternal issues.’

Similarly, VIDs, MVCs, and VPCs were treated as a whole. The major issue we encountered, however, is due to the fact that, unlike NMWEs, VMWEs are highly discontinuous structures leading to issues in annotation, as shown in (20). To overcome this obstacle, layers of annotation provide the dependency graphs that are relative to a sentence. These may be retrieved to account for the structure of the MWE.

- (20) Δεν είναι διαφανής η απόφαση που
 Den ine diafanis i apofasi pou
 Not is.PRS,3SG transparent.SG.NOM the.SG.NOM decision.SG.NOM that
 τελικά έλαβαν.
 telika elavan
 finally take.PST.3PL
 ‘The decision that they finally made was not transparent.’

Discrepancies between the single- and multiword LUs are abundant and need to be identified based on corpus evidence. In the remainder of the section, we will present the mismatches found in our data, which were depicted in the encoding. VMWEs were systematically found to have fewer FEs realised than their single-word counterparts. This is especially true for LVCs as opposed to their single-word counterparts. In most occurrences, the predicative noun is realised

in plural, indicating, thus, an aspectual reading of the LVC, i.e., repetition. In these cases, it is not the verb, but the nominal predicate that triggers the aspectual reading of the whole construction, whereas the verb remains bleached. For example, the multiword LU (el) παίρνω απόφαση.lvc *perno apofasi* (lit. ‘take decision’) ‘to make a decision, to decide’ and the single verb (el) αποφασίζω.v *apofasizo* ‘to decide’ both evoke the Deciding frame defined via the COGNISER and DECISION FES. In our corpus, the LVC at hand was found to systematically realise only the COGNISER FE in the form of a NP in Subject position (in the nominative case), whereas it consistently lacks the DECISION one, as shown in (21); non-core FEs are usually realised as modifiers of the nominal predicate. By contrast, the FE DECISION is realised only in the single word LU as a to-clause, as depicted in (22).

- (21) [Οι ηγέτες_{COGNISER}] παίρνουν [υπεύθυνε_{MANNER}]
 i igetes pernun ypeythines
 the.PL.NOM leader.PL.NOM take.PRS.3PL responsible.PL.ACC
αποφάσεις.
 apofasis
 decision.PL.ACC
 ‘the leaders make decisions in a responsible way.’
- (22) [Ο Γιάννης_{COGNISER}] αποφάσισε [να φύγει_{DECISION}].
 O Gianis apofasise na figi
 The.SG.NOM John.SG.NOM decide.PST.3SG to leave
 ‘John decided to leave.’

Notably, certain VIDs bear a meaning that also incorporates one of their elements, most often intensifiers, but also other arguments as well. In this respect, the VPC in (23) incorporates the FE MANNER that is realised as the adjunct (el) σωστά *sosta* ‘correctly’ assumed by its single word counterpart μαντεύω.v *man-tevo* ‘to guess’. This is due to the fact that the VMWE (el) πέφτω μέσα.vpc *pefto mesa* (lit. ‘fall in’) ‘to guess correctly’, bears a positive reading in contrast to its single-word counterpart (el) μαντεύω.v *madevo* ‘to guess’ that bears a neutral reading. In these cases, we retain the FE at hand in the frame, but we encode it as being realised only in the single-word predicate based on corpus evidence.

- (23) πέφτω μέσα
 pefto mesa
 fall.PRS.1SG in
 ‘to guess correctly’

In most cases, the argument structure of complex predicates deviates from the patterns assumed by their single-word counterparts. This is particularly true about VIDs, due to the fact that they generally follow the valence of their syntactic verb head. For example, the single-word verbal predicate (el) *εξοργίζω.v* *exorgizo* ‘to enrage’ is an Object Experiencer verb, that is, a verb which assumes the FE EXPERIENCER (i.e. the entity that experiences the denoted emotion event); this FE is realised as a NP in the accusative case and in Object position. The CAUSE of the event is realised as an argument, that functions as the Subject of the verb, as shown in (24) (Giouli 2020). In contrast, in the case of the VID (el) *ανεβάζω το αίμα στο κεφάλι* *anevazo to ema sto kefali* (lit. ‘raise.PRS.1SG the.SG.ACC blood.SG.ACC to-the.SG.ACC head.SG.ACC’) ‘to enrage’, the core FE EXPERIENCER is the non-lexicalised element of the VMWE and is realised as a nominal complement (usually, the weak form of the personal pronoun) in the genitive case, whereas the CAUSE of the emotion is realised as a NP in Subject position, as depicted in (25). The weak pronoun (el) *μου* *moy* ‘my’ in the genitive case is due to the valence pattern entailed by the syntactic head of the VMWE; yet, it is annotated as EXPERIENCER.

(24) [O Γιάννης_{CAUSE}] [με_{EXPERIENCER}] εξοργίζει.
 O Giannis me exoryizi
 The.SG.NOM John.SG.NOM me1SG.ACC enrage.PRS.3SG
 ‘John makes me furious.’

(25) [O Γιάννης_{CAUSE}] [μου_{EXPERIENCER}] ανέβασε το
 O Giannis moy anevase to
 The.SG.NOM John.SG.NOM me1SG.GEN raise.PST.3SG the.SG.ACC
αίμα στο κεφάλι.
ema sto kefali
 blood.SG.ACC to.the.SG.ACC head.SG.ACC
 ‘John made me furious.’

Similar discrepancies are attested for other types of MWEs, for example, LVCs. Note that whereas the FE THEME is realised as a NP in the single word LU (el) *αναφέρω.v* *anafero* ‘to mention’, in (26), the same FE is realised as a PP headed by the preposition (el) *σε* *se* ‘to’ in the LVC (el) *κάνω μνεία* *kano mnia* (lit. ‘make.PRS.1SG mention.SG.ACC’) ‘to mention’ as shown in (27). These discrepancies between single- and multiword LUs in the realisation of their FEs have been studied and accounted for in the database based on corpus evidence.

- (26) [Οι Times] αναφέρουν [τις αντιδράσεις_{THEME}].
 I Times anaferoyn tis antidrasis
 The.PL.NOM Times refer.PRS.3PL the.PL.ACC reaction.PL.ACC
 ‘The Times refer to the reactions.’
- (27) [Οι Times] κάνουν μνεία [στις αντιδράσεις_{THEME}].
 I Times kanun mnia stis antidrasis
 The.PL.NOM Times make.PRS.3PL reference.SG.ACC to.the.PL.ACC
 reaction.PL.ACC
 ‘The Times refer to the reactions.’

Finally, syntactic alternations (i.e., passivisation, causative-inchoative alternation, etc.) that are attested for the single-word predicates of a frame are also attested for their VMWE counterparts, yet with different verbs as syntactic heads. This holds true for VIDs and LVCs alike. Indeed, LVCs which comprise the light verbs (el) βγάζω *vgazo* ‘to take out’ and (el) βγαίνω *vgeno* ‘to be taken out’ combined with the same predicative noun signal the causative – inchoative alternation and, in most cases, are assumed under the same frame. They predominately differ in the syntactic function of their lexicalised elements; as a result, the difference between the two is also depicted via their FEs and the grammatical function they assume. For example, the LVCs (el) βγάζω συμπεράσμα.lvc *vgazo symperasma* (lit. ‘take-out.PRS.1SG conclusion.SG.ACC’) ‘to conclude’ and (el) βγαίνει συμπεράσμα.lvc *vgeni symperasma* (lit. ‘is-taken-out.3SG conclusion.SG.NOM’) ‘it is concluded’ enter in the causative-inchoative alternation. In the former, the lexicalised element is the argument in object position (and following the rules of the language, it is realised as a NP in the accusative case); on the contrary, the latter has an argument in subject position as the lexicalised element. They are both assigned in the same Coming-to-Believe frame, yet different FEs are realised for each one of them, since the two multiword LUs differ in the perspective: for the former, the COGNISER is realised, whereas the latter occurs with the THEME as shown in (28) and (29):

- (28) [Οι πολίτες_{COGNISER}] βγάζουν τα συμπεράσματά τους.
 I polites vgazun ta simperasmata tus
 The.PL.NOM citizen.PL.NOM take.out.PRS.3SG the.PL.ACC conclusion.PL.ACC their3SG
 ‘Citizens come to a conclusion.’

- (29) *Βγαίνει* *το* *συμπέρασμα* [*ότι η χώρα*
vgeni *to* *simperasma* *oti i chora*
 go.out.PRS.3SG.PRES the.SG.NOM conclusion.SG.NOM that the country
*κινδυνεύει*_{THEME}].
kindinevi
 is-in-danger
 ‘It is concluded that the country is in danger.’

7 Conclusions

We have presented work aimed at encoding MWEs that pertain to the grammatical categories of noun and verb to a frame-based lexical resource for Modern Greek. The work reported here is part of a larger initiative to construct a lexical database for Modern Greek with an inventory of language-specific frames around which to organise lexical units along the principles already set by BFN and other frame-based resources. Our MWE exploration has also taken into account multiword terms that pertain to the financial domain besides MWEs from the general language. For each MWE, we wish to provide information with respect to frame membership, valence, and access to a large number of annotated examples. Relations with other LUs (both single- and multiword ones) via the frame-to-frame relations already available in the resource have also been defined. The internal structure of the MWEs and their syntactic variations are depicted by means of the annotation layers that are available as pre-processing of the corpus; the focus is no longer on the representation of the internal structure of MWEs and their lexicalised elements, but on their valences; these are depicted via the annotated instances that are included as examples in the database.

Our contribution is two-fold: on the one hand, we provide an overview of the treatment of various types of MWEs in the Greek FrameNet aimed at mapping form onto meaning; on the other hand, we focus on the discrepancies between MWEs and their single-word counterparts. As we have shown, VMWEs were systematically found to have fewer FEs realised than their single-word counterparts bearing the same meaning. Moreover, in LVCs when the predicative noun is realised in plural an aspectual reading of the LVC is possible, i.e., repetition; this aspectual reading is also due to the missing FEs that denote a change in perspective. In a way, this type of representation allows us to provide the deep semantics of MWEs in a way that is comparable to the treatment of single-word lexical entries. For cases of polysemy and near synonymy, the strong apparatus of frame semantics allows us to explore distinct meanings of MWEs that pertain

to LSP (terms) and general language lexical entries alike via frame assignment and FE definition.

The work on FN-el is still in progress, and encoding is continuously subject to refinements and modification. Future work has already been planned towards enriching FN-el with new frames and LUs, both single and multiword ones. In another line of research, the alignment of FN-el frames with the BFN ones is currently underway. Finally, since this lexical resource provides the representation of the lexical and syntactic properties of the MWEs only via the annotated data, we plan to link FN-el to an existing lexical resource for Modern Greek that bears this information.

Abbreviations

BFN	Berkley FrameNet
FE	Frame element
FN-el	Greek FrameNet
LU	Lexical unit
LVC	Light verb construction
MWE	Multiword expression
NMWE	Nominal multiword expression
NP	Noun phrase
PP	Prepositional phrase
VID	Verbal idiomatic expression
VMWE	Verbal multiword expression
VP	Verb phrase
VPC	Verb-particle construction

Acknowledgements

The authors would like to thank the editors and the anonymous reviewers for their comments and insightful suggestions that contributed to improving the manuscript. The research leading to the results presented in this chapter was partially funded by the project “AIO-ILSP: Lexical Resource Infrastructures”, which was financed by the Institute for Language and Speech Processing, ATHENA Research Centre. Corpus annotation and frame assignment were performed by V. Pilitsidou and H. Christopoulos within the framework of the Postgraduate Programme *Translation and Interpreting* of the National and Kapodistrian University of Athens, Faculty of Turkish Studies and Modern Asian Studies.

References

- Anastasiadis-Symeonidis, Anna. 1986. *Η νεολογία στην κοινή νεοελληνική* ('Neology in Modern Greek'). Thessaloniki: Aristotle University of Thessaloniki.
- Anthony, Laurence. 2005. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Proceedings of the International Professional Communication Conference, 2005 (IPCC 2005)*, 729–737. IEEE.
- Baker, Collin F., Charles J. Fillmore & John B. Lowe. 1998. The Berkeley FrameNet project. In *36th annual meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 1, 86–90. Montreal: Association for Computational Linguistics.
- Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 267–292. Boca Raton, FL: CRC Press.
- Boas, Hans C. 2002. Bilingual FrameNet Dictionaries for Machine Translation. In M. González Rodríguez & C. Paz Suárez Araujo (eds.), *Proceedings of the third international conference on Language Resources and Evaluation*, 1364–1371. Las Palmas, Spain: European Language Resources Association (ELRA).
- Bonial, Claire, Julia Bonn, Kathryn Conger, Jena D. Hwang & Martha Palmer. 2014a. Propbank: Semantics of new predicate types. In *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 3013–3019. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Bonial, Claire, Meredith Green, Jenette Preciado & Martha Palmer. 2014b. An approach to *take* multi-word expressions. In *Proceedings of the 10th workshop on multiword expressions (MWE2014)*, 94–98. Gothenburg, Sweden: Association for Computational Linguistics.
- Borin, Lars. 2021. Multiword expressions: A tough typological nut for Swedish FrameNet++. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: Harmonization, integration, method development, and practical language technology applications*, 221–259. Amsterdam: John Benjamins.
- Borin, Lars, Dana Danélls, Markus Forsberg, Dimitrios Kokkinakis & Maria Toporowska Gronostaj. 2010. The past meets the present in Swedish FrameNet++. In *Proceedings of the 14th EURALEX International Congress*, 269–281.
- Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó & Manfred Pinkal. 2009. Using FrameNet for the semantic analysis of German: Annotation, representation and automation. In Hans C. Boas (ed.), *Multilingual*

- FrameNets in computational lexicography: Methods and applications*, 209–244. Berlin, New York: De Gruyter Mouton.
- Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod & Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the third international Conference on Language Resources and Evaluation (LREC'02)*, 1934–1940. Las Palmas, Canary Islands: European Language Resources Association (ELRA).
- Candito, Marie, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël de Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot & Laure Vieu. 2014. Developing a French FrameNet: Methodology and first results. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno & Stelios Piperidis Jan Odiijk (eds.), *Proceedings of the ninth international Conference on Language Resources and Evaluation (LREC'14)*, 1372–1379. Reykjavik, Iceland: European Language Resources Association (ELRA). <https://aclanthology.org/L14-1411/>.
- Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag & Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the third international Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Canary Islands: European Language Resources Association (ELRA).
- Di Fabio, Andrea, Simone Conia & Roberto Navigli. 2019. VerbAtlas: A novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 627–637. Hong Kong, China: Association for Computational Linguistics.
- Dolbey, Andrew, Michael Ellsworth & Jan Scheffczyk. 2006. BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In Olivier Bodenreider (ed.), *Formal biomedical knowledge representation: Proceedings of the second international workshop on Formal Biomedical Knowledge Representation (KR-MED 2006), collocated with the 4th International Conference on Formal Ontology in Information Systems (FOIS-2006)* (CEUR Workshop Proceedings 222). Baltimore: CEUR.
- Erk, Katrin, Andrea Kowalski, Sebastian Padó & Manfred Pinkal. 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for*

- Computational Linguistics (ACL'03)*, 537–544. Sapporo, Japan: Association for Computational Linguistics. DOI: 10.3115/1075096.1075164.
- Faber, Pamela. 2011. The dynamics of specialized knowledge representation: Simulational reconstruction or the perception–action interface. *Terminology* 17(1). 9–29.
- Faber, Pamela. 2015. Frames as a framework for terminology. In Hendrik J. Kockaert & Frieda Steurs (eds.), *Handbook of Terminology*, vol. 1. Amsterdam/Philadelphia: John Benjamins.
- Faber, Pamela & Miriam Buendía Castro. 2014. EcoLexicon. In Andrea Abel, Chiara Vettori & Natascia Ralli (eds.), *Proceedings of the 16th EURALEX international congress*, 601–607. Bolzano, Italy: EURAC Research.
- Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.
- Fillmore, Charles J. 1976. Frame Semantics and the nature of language. *Annals of the New York Academy of Sciences* 280. 20–32.
- Fillmore, Charles J. 1977. Scenes-and-frames semantics. In Antonio Zampolli (ed.), *Linguistic structures processing: Fundamental studies in computer science*, vol. 59 (Fundamental Studies in Computer Science), 55–81. Amsterdam; New York; Oxford: North Holland.
- Fillmore, Charles J. 1982. Frame Semantics. In *Linguistics in the morning calm: Selected Papers from SICOL-1981*, 111–137. Seoul, Korea: Hanshin Publishing Company.
- Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di semantica* 6(2). 222–254.
- Fotopoulou, Aggeliki. 1993. *Une classification des phrases à compléments figés en grec moderne: étude morphosyntaxique des phrases figées*. Université Paris VIII. (Doctoral dissertation).
- Fotopoulou, Aggeliki, Stella Markantonatou & Voula Giouli. 2014. Encoding MWEs in a conceptual lexicon. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 43–47. Gothenburg, Sweden: Association for Computational Linguistics.
- Gavriilidou, Zoe. 2013. NN combinations in Greek. *Journal of Greek Linguistics* 13(1). 5–29.
- Giouli, Voula. 2020. *Το σημασιολογικό πεδίο των συναισθημάτων: Ταξινόμηση των ρημάτων της νέας ελληνικής που δηλώνουν συναίσθημα*. ('The semantic field of emotions: A lexicon-grammar account of Greek verbs denoting emotion. Greek. Athens, Greece: National & Kapodistrian University of Athens. (Doctoral dissertation).

- Giouli, Voula. 2023. A model for representing the semantics of MWEs: From lexical semantics to the semantic annotation of complex predicates. *Frontiers in Artificial Intelligence* 6. DOI: 10.3389/frai.2023.802218.
- Giouli, Voula, Vassiliki Foufi & Aggeliki Fotopoulou. 2019. Annotating Greek VMWEs in running text: A piece of cake or looking for a needle in a haystack? In Maria Chondrogianni, Simon Courtenage, Geoffrey Horrocks, Amalia Arvaniti & Ianthi Tsimpli (eds.), *13th International Conference on Greek Linguistics*, 125–134. University of Westminster, London, UK.
- Giouli, Voula, Vera Pilitsidou & Hephhestion Christopoulos. 2020. Greek within the Global FrameNet Initiative: Challenges and conclusions so far. In *Proceedings of the International FrameNet Workshop 2020: Towards a global, multilingual FrameNet*, 48–55. Marseille, France: European Language Resources Association, (ELRA).
- Grégoire, Nicole. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1). 23–39.
- Gross, Maurice. 1975. *Méthodes en syntaxe: Régime des constructions complétives*. Paris: Hermann.
- Gross, Maurice. 1982. Une classification des phrases « figées » du français. *Revue québécoise de linguistique* 11(2). 36–41.
- Hartmann, Silvana, György Szarvas & Iryna Gurevych. 2012. Mining multiword terms from Wikipedia. In Maria Teresa Paziienza & Armando Stellato (eds.), *Semi-automatic ontology development: Processes and resources*, 226–258. IGI Global.
- Hayoun, Avi & Michael Elhadad. 2016. The Hebrew FrameNet project. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC'16)*, 4341–4347. Portorož, Slovenia: European Language Resources Association (ELRA).
- Holton, David, Peter Mackridge & Irene Philippaki-Warbuton. 1997. *Greek: A comprehensive grammar of the modern language*. London; New York: Routledge.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1. 7–36.
- Kim, Jeong-uk, Younggyun Hahm & Key-Sun Choi. 2016. Korean FrameNet expansion based on projection of Japanese FrameNet. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System demonstrations*, 175–179. Osaka, Japan: The COLING 2016 Organizing Committee.

- Kipper, Karin, Anna Korhonen, Neville Ryant & Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation* 42(1). 21–40.
- Kuiper, Koenraad, Heather McCann, Heidi Quinn, Therese Aitchison & Kees van der Veer. 2003. *SAID*. Tech. rep. Philadelphia. DOI: 10.35111/MSVM-T728.
- Laporte, Éric & Stavroula Voyatzi. 2008. An electronic dictionary of French multiword adverbs. In *Proceedings of the LREC workshop towards a shared task for Multiword Expressions (MWE 2008)*, 31–34.
- Lenci, Alessandro, Martina Johnson & Gabriella Lapesa. 2010. Building an Italian FrameNet through semi-automatic corpus analysis. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC'10)*, 12–19. Valletta, Malta: European Language Resources Association (ELRA).
- Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.
- Lindén, Krister, Heidi Haltia, Juha Luukkonen, Antti Olavi Laine, Henri Roivainen & Niina Väisänen. 2017. FinnFN 1.0: The Finnish frame semantic database. *Nordic Journal of Linguistics* 40(3). 287–311.
- Markantonatou, Stella, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chiril, Dimitrios Karamatskos, Nicolaos Valeontis & George Pavlidis. 2024. Description of Pomak within IDION: Challenges in the representation of verb multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 39–72. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998633.
- Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019) at ACL 2019*, 130–134. Florence. DOI: 10.18653/v1/W19-5115.
- Mini, Marianna. 2009. *Linguistic and psycholinguistic study of fixed verbal expressions with fixed subject in Greek: A morphosyntactic analysis, lexicosemantic gradation and processing by elementary school children*. University of Patras. (Doctoral dissertation).

- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Portorož, Slovenia: European Language Resources Association (ELRA).
- Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.
- Odijk, Jan. 2013. Identification and lexical representation of multiword expressions. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch: Results by the STEVIN programme* (Theory and Applications of Natural Language Processing), 201–217. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-30910-6_12.
- Ohara, Kyoko, S. Fujii, Hiroaki Saito, S. Ishizaki, T. Ohori & Ryoko Suzuki. 2003. The Japanese FrameNet project: A preliminary report. In *Proceedings of Pacific Association for Computational Linguistics (PACLING'03)*, 249–254. Halifax, Canada: Pacific Association for Computational Linguistics.
- Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.
- Ostroški Anić, Ana & Ivana Brač. 2022. Airframe: Mapping the field of aviation through semantic frames. In Annette Klosa-Kückelhaus, Stefan Engelberg, Christine Möhrs & Petra Storjohann (eds.), *Dictionaries and society: Proceedings of the XX EURALEX international congress*, 334–345. Mannheim: IDS-Verlag.
- Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1). 71–106.
- Petruck, Miriam R. L. 1997. Frame semantics. In Jef Verschueren, Jan-Ola Östman, Jan Blommaert & Chris Bulcaen (eds.), *Handbook of pragmatics*, 1–13. Amsterdam: John Benjamins.
- Petruck, Miriam R. L. & Michael Ellsworth. 2016. Representing support verbs in FrameNet. In *Proceedings of the 12th workshop on Multiword Expressions*, 72–77. Berlin: Association for Computational Linguistics.

- Pilitsidou, Vera & Voula Giouli. 2020. Frame Semantics in the specialized domain of finance: Building a termbase to aid translation. In Zoe Gavrilidou, Maria Mitsiaki & Asimakis Fliatouras (eds.), *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*, vol. 1, 263–271. Alexandroupolis: Democritus University of Thrace.
- Ralli, Angela. 2007. *Η Σύνθεση των Λέξεων: Διαγλωσσική, Μορφολογική Προσέγγιση* ('*Compounding: A cross-lingual, morphological approach*'). Athens: Patakis.
- Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archana Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Gungör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya & Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan & Miriam R. L. Petruck (eds.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 222–240. Santa Fe, NM: Association for Computational Linguistics.
- Ramisch, Carlos, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archana Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Gungör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh & Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberiu, Carlos Ramisch, Ashwini Vaidya, Petya Osenova & Agata Savary (eds.), *Proceedings of the joint workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)*, 107–118. Barcelona: Association for Computational Linguistics.
- Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson & Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*. <https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf>.
- Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander F. Gelbukh (ed.), *Proceedings of the third international conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, 1–15. Springer.
- Sager, Juan C. 1990. *A practical course in terminology processing*. Amsterdam: John Benjamins.

- Saito, Hiroaki, Shunta Kuboya, Takaaki Sone, Hayato Tagami & Kyoko Ohara. 2008. The Japanese FrameNet software tools. In *Proceedings of the sixth international conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).
- Salomão, Maria Margarida M. 2009. Framenet Brasil: Um trabalho em progresso. *Caleidoscópio* 7(3). 171–182.
- Savary, Agata, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archana Bhatia, Marie Candito, Polona Gantar, Uxo Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze & Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 24–35. Dubrovnik, Croatia: Association for Computational Linguistics. <https://aclanthology.org/2023.mwe-1.6>.
- Savary, Agata, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxo Iñurrieta & Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics* 112(1). 5–54.
- Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 31–47. Valencia, Spain: Association for Computational Linguistics. DOI: 10.18653/v1/W17-1704.
- Schmidt, Thomas C. 2009. The Kicktionary: A multilingual lexical resource of football language. In Hans C. Boas (ed.), *Multilingual FrameNets in computational lexicography: Methods and applications*, 101–134. Berlin, New York: De Gruyter Mouton.
- Schneider, Nathan, Dirk Hovy, Anders Johannsen & Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th international workshop on Semantic Evaluation (SemEval-2016)*, 546–559. San Diego, California: Association for Computational Linguistics.
- Shudo, Kosho, Akira Kurahone & Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of*

- the Association for Computational Linguistics: Human Language Technologies*, 161–170. Portland, OR: Association for Computational Linguistics.
- Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics.
- Subirats, Carlos. 2009. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In Hans C. Boas (ed.), *Multilingual FrameNets in Computational Lexicography*, 135–162. Berlin/New York: Mouton de Gruyter.
- Tayyar Madabushi, Harish, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th international workshop on Semantic Evaluation (SemEval-2022)*, 107–121. Seattle, WA: Association for Computational Linguistics.
- Timponi Torrent, Tiago & Michael Ellsworth. 2013. Behind the labels: Criteria for defining analytical categories in FrameNet Brasil. *Veredas* 17. 44–65. <https://periodicos.ufjf.br/index.php/veredas/article/view/25403>.
- Timponi Torrent, Tiago, Michael Ellsworth, Collin Baker & Ely Edison da Silva Matos. 2018. The Multilingual FrameNet shared annotation task: A preliminary report. In *Multilingual FrameNets and constructions, The international FrameNet workshop 2018*.
- Timponi Torrent, Tiago, Ely Edison Da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz Da Costa & Mateus Coutinho Marim. 2022. Representing context in FrameNet: A multi-dimensional, multimodal approach. *Frontiers in Psychology* 13. 1–20. DOI: 10.3389/fpsyg.2022.838441.
- Timponi Torrent, Tiago, Maria Margarida M. Salomão, Fernanda C. A. Campos, Regina M. M. Braga, Ely E. S. Matos, Maucha A. Gamonal, Julia A. Gonçalves, Bruno C. P. Souza, Daniela S. Gomes & Simone R. Peron. 2014. Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup. In *Proceedings of COLING 2014, the 25th international conference on Computational Linguistics: System demonstrations*, 10–14. Dublin, Ireland: Dublin City University & Association for Computational Linguistics.
- Venturi, Giulia, Alessandro Lenci, Simonetta Montemagni, Eva Maria Vecchi, Maria-Teresa Sagri, Daniela Tiscornia & Tommaso Agnoloni. 2009. Towards a FrameNet resource for the legal domain. In Núria Casellas, Enrico Francesconi, Rinke Hoekstra & Simonetta Montemagni (eds.), *Proceedings of the 3rd workshop on Legal Ontologies and Artificial Intelligence Techniques, held in conjunction with the 2nd workshop on Semantic Processing of Legal Text* (CEUR Work-

- shop Proceedings 465), 67–76. Barcelona, Spain. <https://ceur-ws.org/Vol-465/paper8.pdf>.
- Villavicencio, Aline, Ann Copestake, Benjamin Waldron & Fabre Lambeau. 2004. Lexical encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating processing*, 80–87. Barcelona, Spain: Association for Computational Linguistics.
- Virk, Shafqat Mumtaz, Dana Dannélls, Lars Borin & Markus Forsberg. 2021. A data-driven semi-automatic framenet development methodology. In *Proceedings of the international conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1471–1479. Held Online: INCOMA.
- Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho & Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics: System demonstrations*, 1–6. Sofia, Bulgaria: Association for Computational Linguistics.
- Yong, Zheng Xin, Patrick D. Watson, Tiago Timponi Torrent, Oliver Czulo & Collin Baker. 2022. Frame shift prediction. In *Proceedings of the thirteenth Language Resources and Evaluation Conference (LREC'13)*, 976–986. Marseille, France: European Language Resources Association (ELRA).
- You, Liping & Kaiying Liu. 2005. Building Chinese FrameNet database. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, 301–306. Wuhan, China: IEEE.
- Zaninello, Andrea & Malvina Nissim. 2010. Creation of lexical resources for a characterisation of multiword expressions in Italian. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC'10)*, 654–661. Valletta, Malta: European Language Resources Association (ELRA).

Chapter 6

Multiword expressions, collocations and the OntoLex vocabulary

👤 Christian Chiarcos^a, 👤 Maxim Ionov^b, 👤 Elena-Simona Apostol^c, 👤 Katerina Gkirtzou^d, 👤 Besim Kabashi^e, 👤 Anas Fahad Khan^f & 👤 Ciprian-Octavian Truică^c

^aApplied Computational Linguistics, University of Augsburg, Germany

^bInstitute for Digital Humanities, University of Cologne, Germany ^cComputer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica

Bucharest ^dInstitute of Language and Speech Processing, Athena Research

Center, Athens, Greece ^eComputational and Corpus Linguistics, University of

Erlangen-Nuremberg, Germany ^fConsiglio Nazionale delle Ricerche - Istituto di Linguistica Computazionale «A. Zampolli», Italy

We describe challenges in and approaches for modelling multiword expressions in machine-readable dictionaries. OntoLex is a widely used community standard for lexical resources on the web, and the predominant RDF vocabulary for the purpose. The current challenge is for OntoLex users to figure out the correct modelling strategy, as different use cases require the application of different OntoLex modules. This chapter serves as an orientation point for researchers and practitioners, and for a number of real-world use cases it will describe modelling strategies and compare their advantages and disadvantages.

1 Introduction

OntoLex (McCrae et al. 2017) is a widely used vocabulary for modelling lexical resources such as lexicons and machine-readable dictionaries on the Semantic

Christian Chiarcos, Maxim Ionov, Elena-Simona Apostol, Katerina Gkirtzou, Besim Kabashi, Anas Fahad Khan & Ciprian-Octavian Truică. 2024. Multiword expressions, collocations and the OntoLex vocabulary. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 187–227. Berlin: Language Science Press. DOI: 10.5281/zenodo.



Web as Linguistic Linked (Open) Data (LL(O)D).¹ It is worth noting, however, that OntoLex was not originally designed as a vocabulary for publishing language resources per se; instead it was developed, at least initially (that is, during the drafting of its original modules) for the rather more specialised task of ontology lexicalisation. Unsurprisingly, this resulted in design decisions (again, at least in its original modules) that were and that remain relatively nontransparent to many linguists, lexicographers and Natural Language Processing (NLP) engineers; with many of these design decisions pertaining to OntoLex's treatment of multiword expressions (MWEs). Our aim, therefore, in the following chapter is to provide detailed orientation as to which of the modelling options offered by OntoLex are most appropriate for describing the most salient aspects of multiword expressions. We consider this to be a necessary contribution at this point in time as there are several alternative modelling options for encoding individual aspects of MWEs within OntoLex, each with their specific characteristics, benefits and downsides. However, before diving too far into the details of OntoLex, we will begin by clarifying what we understand by *multiword expressions* in the rest of this chapter, and what we view as being the primary modelling needs and requirements in relation to such kinds of linguistic phenomena.

1.1 Background: Multiword expressions

We define MWEs as linguistic forms that span conventional word boundaries and, following Sag et al. (2002), we also define them as combinations of words for which the semantic or syntactic properties of the entire expression cannot be predicted from its parts. This is generally compatible with the view on MWEs and collocations taken by other theoretical frameworks, e.g., Meaning-Text Theory, which views them as linguistic units that consist of two or more words functioning as a single semantic and syntactic entity (Mel'čuk 2006). According to Hüning & Schlücker (2015), the main types of MWEs include the following: idioms (*to kick the bucket*), metaphors (*as sure as eggs is eggs*), stereotyped comparisons (*swear like a trooper*), proverbs (*A bird in the hand is worth two in the bush*), quotations (*shaken, not stirred*), commonplaces (*one never knows*), binomial expressions (*shoulder to shoulder*), complex nominals (*weapons of mass destruction*), syntactic noun incorporation ((de) *Auto waschen* 'to car wash'), particle verb constructions (*to make up*), complex predicates (*to have a look*), fossilized forms (*all*

¹The specifications for OntoLex can be consulted at <https://www.w3.org/2016/05/ontolex/>. If you wish to participate in the development of future OntoLex modules, please join the W3C Ontology Lexicon group <https://www.w3.org/community/ontolex/>. In addition, you can raise issues about the vocabulary at the OntoLex GitHub <https://github.com/ontolex/>.

of a sudden), routine formulas (*Good morning*), and collocations (cf. Evert 2005, 2009, Schlücker 2019, Finkbeiner & Schlücker 2019).

Note that Hüning and Schlücker's use of the term collocation here is somewhat ambiguous in that they seemingly refer to the (more limited) case of *lexicalized* collocations, namely, those collocations that exhibit non-compositional semantics or lexical selection preferences: e.g., the phrase *brush one's teeth* is a common expression in English, whereas *polish one's teeth* or *wash one's teeth* are not. However, in corpus linguistics, the term collocation refers to *any* set of words whose likelihood of co-occurrence is greater than a certain pre-determined threshold figure as determined by salient collocation metrics; this is also how we will understand collocations in the rest of the chapter. On this account, not every collocation observed in a corpus is a MWE, but lexicalised collocations and other MWEs generally exhibit high collocation scores, so automated collocation analysis can also be used for lexicographic purposes.

Indeed, OntoLex was developed to take into account the functionality of several tools developed for such (lexicographically oriented) purposes, e.g., Sketch Engine (Kilgarriff et al. 2014), Corpus WorkBench² (Evert & Hardie 2011) and CQPweb (Hardie 2012) – so that even if these tools do not have machine-readable interface specifications, their APIs are widely used in digital lexicography. One of the individual OntoLex modules which we will be discussing below, FrAC (Chiaros et al. 2022a), was specifically designed to address this issue and follows the requirements of these and other tools (as well as taking into consideration several other aspects of corpus-based information in lexical resources). But FrAC is not the only part of the OntoLex vocabulary that is relevant to the modelling of MWEs. However, in order to clarify this statement, it will be necessary to anticipate the more detailed analysis of OntoLex offered later in this chapter and give a brief resume of how the vocabulary is structured and see how it can be used to describe MWEs.

1.2 Background: Describing MWEs with Linguistic Linked Data

The OntoLex vocabulary consists of a number of modules, four of which were part of the original specifications published in 2016. These include a core module (OntoLex-Core), along with modules dealing with: *syntax and semantics* and in particular syntactic and semantic frames (*synsem*);³ the *decomposition* of MWEs

²<https://cwb.sourceforge.io/>

³<https://www.w3.org/2016/05/ontolex/#syntax-and-semantics-synsem>

and compounds (**decomp**);⁴ *variation and translation* (**vartrans**);⁵ and linguistic metadata (**lime**).⁶ A further module dealing with lexicographic use cases (**lexicog**) was published in 2019 as part of a subsequent W3C Community Report,⁷ and two new modules **FrAC** and **morph** are currently in advanced stages of development and will be further described in Sections 3.2 and 3.3, respectively.

In terms of a brief summary of the provision offered by these various different OntoLex modules for modelling multiword expressions and compound words,⁸ we can say the following: **OntoLex-Core** (Sect. 2.1) introduces the concept `ontolex:MultiWordExpression` as a subclass of `LexicalEntry`; **decomp** offers a model to describe the *inner structure* of multiword expressions (McCrae et al. 2016); **FrAC** addresses metrics, techniques and data structures for automatically identifying *collocations in corpora*, for compiling of *collocation dictionaries* and for the linking of dictionaries with *attestations of MWEs (qua lexical entries)* in corpora (Chiarcos et al. 2022a,c); finally, morphological compounding is a morphological process that in some languages (e.g., German and English) creates multiword expressions, and morphological aspects of MWEs are consequently addressed by the emerging **morph** module dealing with morphology (Chiarcos et al. 2022d).

The distribution of these different aspects of the modelling or description of MWEs across four different OntoLex modules (**OntoLex-Core**, **decomp**, **FrAC** and **morph**) may cause misunderstandings or uncertainties as to which strategy should be used for which particular type of resource or use case. At the very least, there is a risk that people looking for ways to model multiword expressions in OntoLex will stop searching as soon as they encounter `ontolex:MultiWordExpression` in the **OntoLex-Core** module. This may not be incorrect in many cases, but it might not be the best solution under all circumstances.

Aside from discussing the details of the provision offered by OntoLex for modelling MWE data (the *how*), another goal of this chapter is to demonstrate the applicability and advantages of doing this in the first place (the *why*). We therefore posit the following requirements for modelling (lexical resources containing) multiword expressions or collocations: namely, a vocabulary for MWEs on the web should support:

⁴<https://www.w3.org/2016/05/ontolex/#decomposition-decomp>

⁵<https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans>

⁶<https://www.w3.org/2016/05/ontolex/#metadata-lime>

⁷<https://www.w3.org/2019/09/lexicog/>

⁸Note here that we are once again anticipating topics which will be described in greater detail in the rest of the chapter.

- the *identification* or categorisation of MWEs as a special type of lexical entry, in order to be able to describe their specific senses and distinguish them from non-lexicalized phrasal expressions,
- *different structural analyses* thus allowing the description of MWEs *either* as opaque units *or* by providing an analysis of their internal structure,
- the provision of *collocation scores* to represent candidate MWEs *together with* a numerical assessment of their likelihood,
- *dynamic prediction* to permit the encoding of the output of web services and automated tools that produce such analyses from corpora, and
- *extensibility and customizability* to allow for the provision of usage examples, and detailed, resource-specific metadata or analyses.

In terms of resource types covered, a vocabulary for MWEs and for the analysis of MWEs should take into consideration legacy resources for multiword expressions, idiomatic expressions and collocations, including, but not limited to classical print dictionaries, dedicated collocation dictionaries, or portals and tools for corpus-based lexicography. At the same time, it should be equally applicable to web services that provide established methods for corpus analysis.

2 The OntoLex Vocabulary

The web of data is grounded on standards such as HTTP, URIs, and RDF; these enable the effortless linking of, and information aggregation over, distributed data on the web. RDF technologies have been widely adopted for linguistic data and machine-readable dictionaries, thanks in particular to their enabling of transitive querying across multilingual lexical resources such as dictionaries and their seamless integration of linguistic resources with either knowledge graphs (ontologies and term bases) or electronic text (corpora and data streams).

OntoLex is the dominant community standard for this kind of data, and its development was guided by five key principles: (1) it should be an RDF model with OWL semantics (Bechhofer et al. 2004), (2) it should support multilinguality and avoid language-specific biases, (3) it should provide semantics by reference vis-à-vis external vocabularies, (4) it should be open, with no costs or licensing restrictions and allow contributions from any and all interested parties, and (5) it should reuse relevant standards and models wherever appropriate. As we have

already stated, **OntoLex** consists of several modules. The core module, **OntoLex-Core**, originates from an earlier RDF vocabulary (McCrae et al. 2010), which was developed on the basis of LexInfo (Cimiano et al. 2011) and LMF (Francopoulo et al. 2009). Since 2011, **OntoLex** has been developed and maintained by the W3C Ontology-Lexica Community Group. Moreover, since the publication of the core vocabulary in 2016, the community group has continued to develop new **OntoLex** modules with an eye to increasing the practicality and versatility of the model and to ensuring its applicability to the needs of further groups of users and types of resources.

2.1 **OntoLex-Core** and **OntoLex** Modules

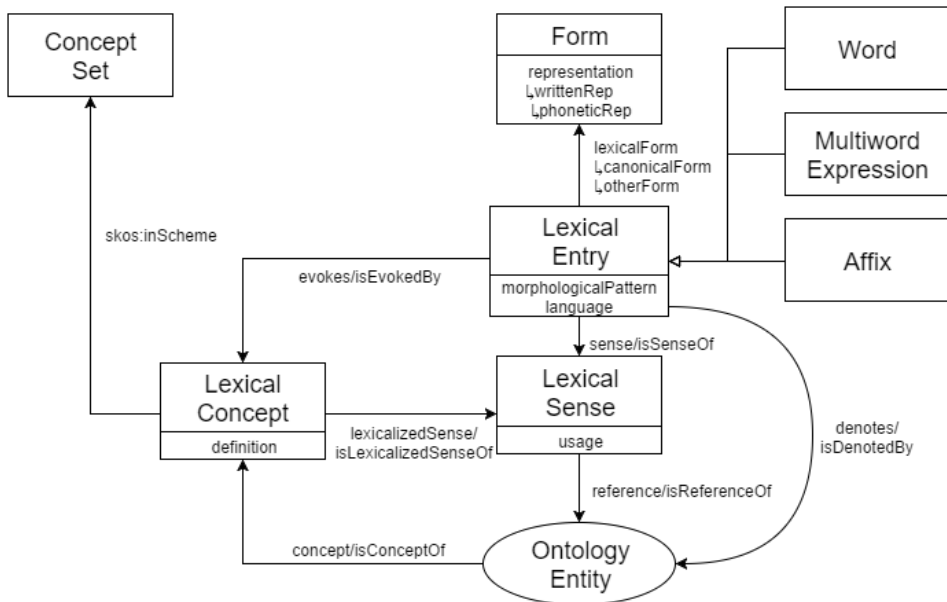


Figure 1: **OntoLex-Core**.

OntoLex-Core⁹ (Figure 1) was developed around the notion of `ontolex:LexicalEntry` as the primary unit of analysis/description of a lexical resource. Each `LexicalEntry` is associated with a set of grammatically related forms as well as a set of word senses and related concepts (that is, at least from the point of view of the **OntoLex-Core** module, other kinds of linguistic description are provided by additional **OntoLex** modules). The `ontolex:Form` class represents

⁹<https://www.w3.org/2016/05/ontolex/>

one grammatical realisation of a lexical entry, e.g. its written representation, annotated with morphological features, while the `ontolex:LexicalSense` represents one lexical meaning of a lexical entry, e.g., a classical word sense. The `ontolex:LexicalConcept` class is an abstraction over a collection of lexical senses, e.g., a semantic frame, a set of synonyms or a term that can be lexicalised in different ways. This latter class also represents semantic meanings, but differs from senses in being more abstract: lexical concepts can typically be realised by different lexical entries. This distinguishes them from senses which are associated with exactly one lexical entry in the OntoLex model.

Within **OntoLex-Core**, `ontolex:MultiwordExpression` is a subclass of `ontolex:LexicalEntry` and is used to classify lexical entries that consist of two or more words. The core module does not provide vocabulary for further elucidating the internal structure of a MWE,¹⁰ it only allows users to indicate that a lexical entry is a MWE and to provide form and sense information as with any other lexical entry. However, as mentioned above, in addition to the core model, four other OntoLex modules were published in 2016 and in the following section, we will describe **decomp**, the most relevant of these for the current discussion on modelling MWEs. Additionally, in 2019, a novel Lexicography Module, **lexicog** (Bosque-Gil & Gracia 2019), was published to address the representation of traditional print dictionary forms. To prevent information loss in the migration of lexical data to OntoLex, **lexicog** introduces the class `lexicog:Entry` to group together lexical entries and associate shared information, e.g., to replicate the grouping of multiple lexemes under a common head word in a dictionary. Its superclass `lexicog:LexicographicComponent` provides a similar function for sub-entries, lexical senses, lexical forms, etc. For reasons of space, we will not discuss this module further here. Other subsequent extensions include the emerging modules **FrAC** for frequency, attestation and corpus-based information in lexical resources, and **morph**, for morphology. Both are described with further detail below as they are relevant for the current discussion on MWEs.

2.2 Decomposition: **decomp**

The OntoLex decomposition module, namely **decomp** (Figure 2), allows for a formal description of the process of constituting multiword expressions or compound lexical entries. It models decomposition primarily by means of

¹⁰In addition to the internal structure of a MWE, information about the valency of MWEs is also useful. At the time of writing, the provision for modelling of valency information for complex predicates within the OntoLex family of modules is still very much under development. We intend to present further updates on this theme in upcoming work.

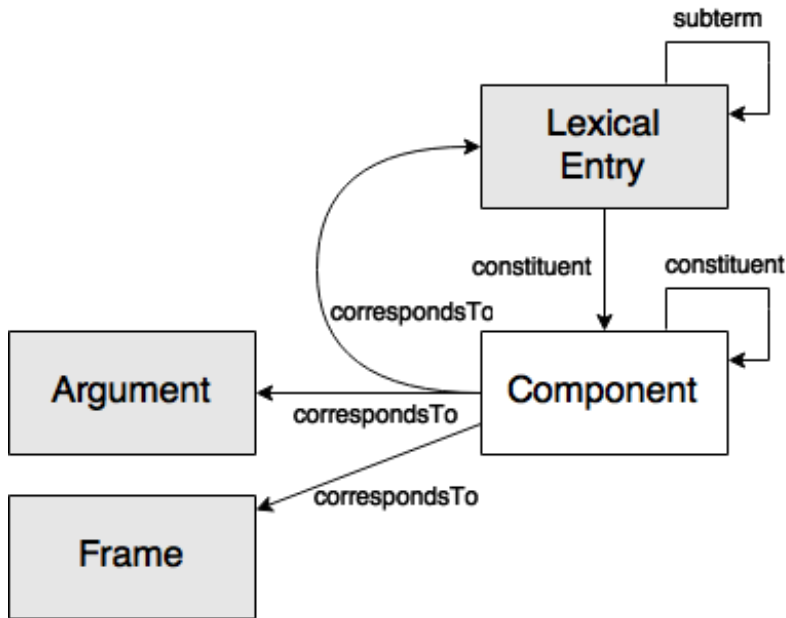


Figure 2: The OntoLex decomp module.

`decomp`: Component, which must uniquely correspond to a lexical entry, a semantic frame or a syntactic argument. Each lexical entry which has been so decomposed then consists of a number of constituents, which correspond to its components, e.g., the division of a nominal compound or a MWE into smaller units. These components can be annotated with morphosyntactic information, such as part of speech or morphological features, and their order can be indicated by `rdf:_n` properties. As a shorthand, lexicons that do not need to represent individual components can use the property `decomp:subterm`.

Aside from basic decomposition, `decomp` allows us to align the sub-units of a composite term with a grammatical role (`synsem:Argument`) or a semantic role (`synsem:Frame`). With `decomp`, we can thus express both the semantics of a phrase and the semantics of the individual lexemes, and beyond that, we can express the semantic relations between these terms in a specific multiword expression by mapping syntactic relations that hold between them and semantic frames (for an idea of how syntactic information might be aligned with information relating to the decomposition of a MWE in `decomp` see the *to know* example in the W3C OntoLex guidelines).¹¹ Frames are defined by the `synsem` module and not

¹¹<https://www.w3.org/2016/05/ontolex/#phrase-structure>

further discussed here, the important aspect is, however, that **decomp** provides the necessary means to represent (a) the lexical semantics of the respective components, (b) the semantics of the MWE as a whole, and (c) the semantics and syntactic structure of a MWE side-by-side.

2.3 Corpus information: OntoLex-FrAC

OntoLex-FrAC (Figure 3) (Chiarcos et al. 2022a) is an emerging vocabulary for enriching machine-readable dictionaries with corpus-based information, relating to word frequency and attestations (Chiarcos et al. 2020), embeddings and distributional similarity (Chiarcos et al. 2021) and collocations (Chiarcos et al. 2022a,c). The core element of FrAC is `frac:Observable`, which refers to anything that can be observed within a corpus, such as forms (`ontolex:Form`), lexemes (`ontolex:LexicalEntry`), but also lexical or ontological concepts, in case this information is present in the data.¹² This definition of observables is organically applicable to collocations, as well.

In FrAC, collocations are not considered as lexical units, but rather as an arbitrary co-occurring group of observables characterised by a collocation score. Since collocations can consist of two or more words, we model `frac:Collocation` as an RDF container of `frac:Observables`, not as a relationship between words. Also, collocations themselves are taken to be `frac:Observable` entities, possessing properties such as attestations, frequency information, similarity scores, etc. Additional parameters, such as the size of the context window used for collocation analysis can be provided in human-readable form in `dct:description`.

In automated collocation analysis, collocations can be described with various collocation scores (`frac:cscore`, sub-property of `rdf:value`). If multiple metrics are used, then the appropriate sub-property of `frac:cscore` should be used.¹³ For asymmetric scores (e.g., relative frequency, `frac:relFreq`), we distinguish the lexical element they are about (using the property `frac:head`) from its collocate(s).¹⁴

¹²This enumeration is vague by design since we expect that other classes that define various corpus annotations (within or outside of OntoLex) could be defined as subclasses.

¹³For specific collocation metrics within FrAC see Appendix A.

¹⁴The property `frac:head` is restricted to indicate the directionality of asymmetric collocation scores. It must not be confused with the notion of *head* in certain fields of linguistics, e.g., in dependency syntax or morphological compounding. Also, it should not be used to model the structure of collocation dictionaries into headwords and associated collocations – for this function, please resort to *lexicog*.

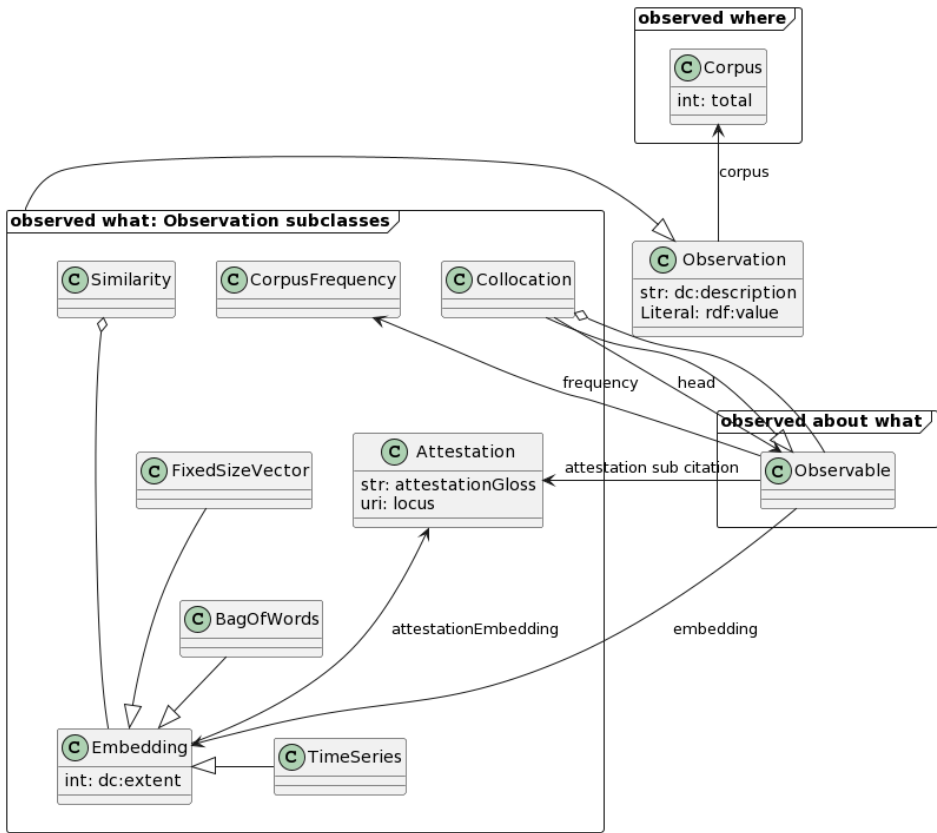


Figure 3: The OntoLex-FrAC module as an UML class diagram (see Suchánek & Pergl (2020) for notation), version July 2022.

2.4 Morphology: OntoLex-Morph

The OntoLex-Morph module is an emerging module designed for describing *both* the morphological structure of linguistic forms/lexical entries) in morphological dictionaries (Klimek et al. 2019) *and* the processes and technical components for generating and parsing inflected or derived word forms as used in computational applications (Chiarcos et al. 2022d).

The class `morph:Morph` is a subclass of `ontoLex:LexicalEntry` that represents a concrete primitive element of (morphological) analysis. An OntoLex morph is like a morpheme in that it constitutes a lexical entry, i.e., a lexicalised or grammaticalised morphological unit, but at the same time, it differs from the classical understanding of *morpheme* in that different allomorphs of the same morpheme can be modelled as distinct morphs – if needed.

of other OntoLex modules. The overall goal of the current section, then, is to delineate strategies for combining and/or choosing between **decomp**, **morph** or **FrAC**, on the basis of the intended use case. Generally speaking, **decomp** deals with the internal structure and combinatory semantics of MWEs, whereas **morph** deals with their morphological structures. **FrAC** deals with collocation analysis, its interplay with MWEs and is described in the following section. Before going into details, however, it should be noted that whereas **morph** and **FrAC** contain relatively little overlap between them, **decomp** has potential overlaps with both **morph** and **FrAC**.

decomp vs. morph: MWEs that involve specialised morphemes (e.g., linking elements that can be used to form nominal compounds) can be described either with **decomp** (in case the resource or task calls for an emphasis on their semantics), with **morph** (in case the resource or task calls for an emphasis on their morphology), or with elements from both vocabularies, depending on the situation in question. The intention is that **decomp** should be used in cases in which we wish to give a “shallow” morphological description of a MWE; it should therefore be considered the default choice and will be suitable for most non-specialist use cases. Alternatively, **morph** (optionally in conjunction with **decomp**) to be preferred in cases where a more “in-depth” morphological description of MWEs, and their constituents, is to be given: namely, where the focus is on the analysis of individual morphemes.

decomp vs. FrAC: **Decomp** and **FrAC** offer two opposing strategies for the analysis of MWEs/collocations – top-down and bottom-up, respectively. **Decomp** provides a mechanism for splitting a lexical entry into smaller components, whereas **FrAC** collocations consist of several observables (e.g. lexical entries). Due to this, **decomp** is preferred for collocations and MWEs that are *confirmed* lexical entries (with optional **FrAC** collocation scores), such as idiomatic expressions, and the emphasis is on their metadata. On the other hand, the **FrAC** collocation class should be used primarily for cases in which the emphasis is on the collocations and their components, especially if they are represented in a corpus or extracted from there by automated methods. Additionally, **FrAC** should be used for collocations with variable word order since **decomp** requires fixed order of the components and **FrAC** only requires observables to occur in the same context (even if they have other words in between).

3.1 *OntoLex-Core*: Declaring a lexicalized multiword expression

MWEs that are confirmed as lexical entries in their own right can be represented as individuals of `ontolex:MultiWordExpression` class; sense information may then be associated with individual such MWEs via the `ontolex:sense` property. The `LexInfo` property `lexinfo:termType` can be used to give a more fine-grained classification of these MWEs as e.g., one of `lexinfo:compound`, `lexinfo:idiom`, `lexinfo:phraseologicalUnit` or `lexinfo:setPhrase`. In addition, the `FrAC` module can be used to describe the frequency and distribution of a MWE in a corpus and provide evidence of its status as a lexical unit.

We illustrate this with the word *cat's-eye*, *cat's eye* or *catseye* by which is meant a retroreflective safety device used in road markings.¹⁵ In this case, we assume that we are dealing with a multiword expression with different orthographic variants. Using the *OntoLex-Core* vocabulary, we can state that it is a (lexicalised) MWE with its specific meaning:¹⁶

```
:cat_s_eye_lex a ontolex:LexicalEntry, ontolex:MultiwordExpression ;
  ontolex:canonicalForm
    [ ontolex:writtenRep "cat's eye"@en, "cat's-eye"@en, "catseye"@en ] ;
  ontolex:sense
    [ ontolex:reference <http://dbpedia.org/resource/Cat's_eye_(road)> ] .
```

Of course, separate lexical entries for `:cat` and `:eye` can be added, but we need specialised modules to clarify their relationship.¹⁷

3.2 *decomp*: MWE Syntax and Semantics

We decompose the entry into its constituent terms `:cat_lex` and `:eye_lex` (each an *OntoLex* lexical entry in its own right):

```
:cat_s_eye_lex decomp:subterm :cat_lex ; decomp:subterm :eye_lex .
```

¹⁵We broadly follow Wiktionary (<https://en.wiktionary.org/wiki/cat's-eye>), but also cf. *cat's eye* in Brewer et al. (1991), and *catseye* in the Longman Dictionary of Contemporary English, <https://www.ldoceonline.com/dictionary/catseye>.

¹⁶Note that in the following listing and in the rest of this chapter we will be using the turtle syntax, see <https://www.w3.org/TR/turtle/>.

¹⁷We exclude the *lexicog* vocabulary here. It is, indeed, capable of expressing the *placement* of the phrase *cat's eye* under the head word *cat* (as in Brewer et al. 1991: 88), but this carries no information about the function and meaning of this grouping preference. For this, we need *decomp*, *morph* or *FrAC* in addition to *lexicog*.

According to the OntoLex specifications, “[i]t is important to mention that the subterm property is a relation between lexical entries and neither indicates the specific inflected word of a lexical entry that appears in the compound nor the position at which it appears”.¹⁸ The structure of the entry does not thus fully reflect the surface strings. Also, in this example, the genitive morpheme ’s is not expressed in the decomposition – neither in **OntoLex-Core** nor in **decomp**, would we normally consider this a lexical entry in its own right.

Alternatively, in **decomp**, we can use the Component class to reflect the particular realisation of a lexical entry that forms part of a compound lexical entry:

```
:cat_s_eye_lex decomp:constituent :cat_s_const ; decomp:subterm :eye_lex .  
:cat_s_const a decomp:Component ; decomp:correspondsTo :cat_lex .
```

Optionally, morphosyntactic constraints can be added to a component. As an example, the string *cat’s* (resp. *cats-* in *catseye*) can be interpreted as a genitive singular. This analysis can be added to `:cat_s_const`:

```
:cat_s_const lexinfo:number lexinfo:singular ;  
lexinfo:case lexinfo:genitive .
```

This analysis captures the syntactic (constituent) structure of the MWE, and it is assumed to be unique. In addition to that, a semantic interpretation can be given by creating `decomp:correspondsTo` relations between a `decomp` component and a `synsem:Argument` or a `synsem:Frame`. We now model the same example using `morph` and highlight the differences in the kinds of information which can be expressed.

3.3 OntoLex-Morph: MWE morphology

Languages differ in the extent to which they employ morphology in the formation of multiword expressions. In English, this is relatively rare, but exhibited in our example. The modelling of *cat’s eye* above did not require the use of the **morph** vocabulary. Indeed, we suggest using the latter only in case a detailed analysis at the level of individual morphemes is required. This is not necessary in order to simply point out that *cat’s* is a genitive form (this can be a morphosyntactic feature of the component) but *is* necessary if we want to provide morpheme-level segmentation, i.e. if we want to state that ’s is a nominal inflection morpheme that indicates genitive singular. For this purpose, **morph** makes use of `morph:Morph`:

¹⁸<https://www.w3.org/2016/05/ontolex/#decomposition-decomp>

6 Multiword expressions, collocations and the OntoLex vocabulary

```
:_s_morph a morph:Morph;  
  ontolex:canonicalForm [ ontolex:writtenRep "'s"@en ] ;  
  morph:grammaticalMeaning  
    [ lexinfo:number lexinfo:singular ; lexinfo:case lexinfo:genitive ] ;  
  morph:baseConstraint [ lexinfo:noun ] .
```

As morph morphs are OntoLex lexical entries, `:_s_morph` could just be added as a `decomp:subterm` as before. A more transparent analysis is to make explicit that it operates as a linking element in a compound:¹⁹

```
:_s_compound_rule a morph:CompoundingRule ;  
  morph:generates :cat_s_eye_lex ; morph:involves :_s_morph .
```

With `morph:replacement`, we can provide one or more different replacement patterns for the morpheme, using standard regular expressions with capturing groups as provided, for example, by the RDF query language SPARQL²⁰ and all major programming languages since Perl:²¹

```
:_s_compound_rule morph:replacement  
  [ morph:source "([s])$" ; morph:target "\\1's" ] .
```

Even without further addenda, these statements can be used to complement the `decomp` analyses given above, as they all refer to the same URI `:cat_s_eye_lex`, each adding more information. Furthermore, **morph** also allows us to add more information about the structure of the compound. For example, we can define a `morph:CompoundHead` relation between the two lexical entries to identify the morphological head of the compound:

```
[ a morph:CompoundHead ;  
  vartrans:source :eye_lex ; vartrans:target :cat_s_eye_lex ] .
```

¹⁹Although this analysis is normally not applied to English, it is the standard way of describing linking morphemes in languages where genitive morphemes in compounds bleached and were subsequently stripped off their original grammatical meaning. German *Katzenauge* (lit. ‘cats’ eyes’) ‘cats’ eye’, uses the linking element *-en-*, originally for a genitive plural. Yet, there is no plural semantics involved: One eye can belong to no more than one cat. Especially with the spelling *catseye*, this way of modelling is appropriate for English as well, as the spelling obfuscates the original genitive marker in a similar way.

²⁰<https://www.w3.org/TR/rdf-sparql-query/#funcex-regex>

²¹Note that this rule describes only one of the three aforementioned orthographic variants, ‘cat’s [eye]’ since every rule should generate exactly one form. To model the other two, additional (alternative) compounding rules must be provided.

In order to link the part of the expression that undergoes morphological transformations with the corresponding rule, we can use a `morph:CompoundRelation`:

```
[ a morph:CompoundRelation ;  
  vartrans:source :cat_lex ; vartrans:target :cat_s_eye_lex ;  
  morph:wordFormationRule :_s_compound_rule ] .
```

Morph word formation relations like `morph:CompoundHead` and `morph:CompoundRelation` are lexical relations as defined in `vartrans`, but in the context of **morph**, they are also reifications of `decomp:subterm` and can be used to provide additional metadata to subterm relations. We use this here to associate a word formation rule with *cat*'s. (Note that we point to the word formation rule only from the node that undergoes morphological transformation modifier because it is the only node that is affected by that replacement.)

In this example, morpheme order is left implicit. However, in concrete applications, it can be inferred from language-specific constraints on the placement of heads and modifiers in morphological compounds.

Note that the reified representation is not the only way to indicate the order of head, modifier, and linking morpheme within a compound. As recommended in **decomp**, the RDF properties `rdf:_1`, `rdf:_2`, etc. can be used to make the order of components explicit. Alternatively, as recommended in **morph**, ordering information can be captured at the level of `ontolex:Form`:

```
:cat_s_eye_lex ontolex:canonicalForm :cat_s_eye_form .  
:cat_s_eye_form a ontolex:Form ;  
  ontolex:writtenRep "cat's eye"@en ;  
  morph:consistsOf :cat_stem, :_s_morph, :eye_stem .  
  rdf:_1 :cat_stem ; rdf:_2 :_s_morph ; rdf:_3 :eye_stem .
```

In this analysis, we introduce separate URIs for the *cat* and *eye* morphemes for the sake of clarity. Alternatively, we can also directly make use of `:cat_lex` and `:eye_lex`, but note that their use as objects of `morph:consistsOf` entails (by RDFS semantics) that these are `morph:Morph` (in addition to the explicitly stated information that they are `OntoLex` lexical entries).

4 Modelling collocations in OntoLex

So far, we have focused on representative lexical examples for illustrating modelling choices. For collocation analysis in **FrAC**, we will need to ground our discussion in real-world data. For reasons of presentation, we focus on relatively simple data, but **FrAC** is equally applicable to more advanced use cases.

4.1 Collocations in OntoLex-FrAC

N-Grams are the most elementary assessment of collocations, and can thus be used for the automatically supported detection of MWEs. *N*-Gram databases are thus practically relevant addenda to lexical resources, but they are normally not seen as full-fledged lexical resources in their own right. In particular, without further analysis, *n*-grams are not necessarily lexicalized MWEs or the result of a morphological process, so they are clearly within the realm of FrAC, and should not be modelled as `ontolex:MultiWordExpression` or by means of `morph` or `decomp`.

A seminal collection of *n*-grams is provided by Google Books²² and features *n*-gram frequencies per publication year as tab-separated values. For example, if we are interested in word usage in the year 2008, the second edition of Google Books provides token and document frequencies for the bigram *cat's + eye*:²³

ngram	year	match_count	volume_count
eye_NOUN	2008	1837106	167735
eyes_NOUN	2008	5672681	176942
cat_NOUN 's_PRT eye_NOUN	2008	515	356
cat_NOUN 's_PRT eyes_NOUN	2008	937	751
cats_NOUN '_PRT eye_NOUN	2008	2	2
cats_NOUN '_PRT eyes_NOUN	2008	169	140

where `match_count` denotes how many times the *n*-gram occurred overall, i.e. *n*-gram frequency, while `volume_count` denotes in how many distinct books of the Google corpus, i.e. document frequency. Note that Google Books provide information about wordforms, not lexemes, so we need to take into account all possible forms of a word in question. On the basis of this, we create OntoLex lexical entries:

```
gb:eye_lex a ontolex:LexicalEntry; lexinfo:partOfSpeech lexinfo:noun;
  ontolex:canonicalForm [ ontolex:writtenRep "eye"@en ] .
```

Since in this example we are interested in a specific time frame only, we can introduce specialised subclasses for collocation and frequency type for this particular corpus and time frame. This is an efficient way to provide a much more compact encoding, as metadata does not have to be repeated for each individual observable.

²²<http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>

²³`eye_NOUN` is retrieved from the file of the English 1-gram (`googlebooks-eng-all-1gram-20120701-e.gz`), while *cat's eye* corresponds to a trigram `cat_NOUN 's_PRT eye_NOUN` and is retrieved from the corresponding list of 3-grams (`googlebooks-eng-all-3gram-20120701-ca.gz`).

```
gb:GB_2008 a owl:Class; # an auxiliary class introduced
  rdfs:subClassOf      # for the convenient handling
    [ owl:Restriction; # of frac:corpus and dct:temporal
      owl:onProperty frac:corpus ;
      owl:hasValue
        <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html> ];
    [ owl:Restriction;
      owl:onProperty dct:temporal; owl:hasValue "2008"^^xsd:date ] .

gb:GB_2008_coll rdfs:subClassOf
  frac:Collocation, frac:Seq, # a class for ordered collocations
  gb:GB_2008 . # that inherits frac:corpus and dct:temporal

gb:GB_2008_doc_freq rdfs:subClassOf
  frac:Frequency, # a frequency class
  gb:GB_2008, # that inherits frac:corpus and dct:temporal
  [ owl:Restriction; # and provides document frequencies
    owl:onProperty dct:description; owl:hasValue "document frequency" ] .

gb:GB_2008_freq rdfs:subClassOf
  frac:Frequency, # a frequency class
  gb:GB_2008, # that inherits frac:corpus and dct:temporal
  [ owl:Restriction; # and provides token frequencies
    owl:onProperty dct:description; owl:hasValue "token frequency" ] .
```

With these corpus-specific classes, we can now provide raw and document frequencies for observables (lexical entries and collocations), as well as relative frequencies (frac:relFreq, obtained from the bigram token frequency divided by the token frequency of the head of the collocation):

```
# unigram (lexeme) frequencies
gb:eye_lex frac:frequency
  [ rdf:value "344677"; a gb:GB_2008_doc_freq ] ,
  [ rdf:value "7509787"; a gb:GB_2008_freq ] .

# bigram (collocation) frequencies
[ rdf:1_gb:cat_lex; rdf:2_gb:eye_lex ] a gb:GB_2008_coll ;
  frac:frequency
    [ rdf:value "1249"; a gb:GB_2008_doc_freq ] ,
    [ rdf:value "1623"; a gb:GB_2008_freq ] ;
  frac:relFreq "0.00022"; # = 1623/7509787
  frac:head gb:eye_lex .
```

The value of `frac:relFreq` corresponds to $p(\langle :cat_lex, :eye_lex \rangle | :eye_lex)$. This can be compared with the relative frequency of `:cat_lex` in the overall corpus to assess its lexicographic significance, calculated from the absolute frequency of lexical entries divided by the `frac:total` number of tokens of the corpus.

This encoding not only provides well-defined datatypes for the information in the original table, but it is also relatively compact: for each bigram in the original database, we produce 3 triples to define components and type, 3 triples per frequency count and type, and 2 triples per collocation score.

4.2 The OZDIC collocation dictionary

The OzDictionary website (OZDIC)²⁴ is a collocation dictionary designed as a learning tool for assisting students in preparing for the Test for English as a foreign language (TOEFL) and similar writing tests. For each headword, the dictionary shows which words and phrases are commonly used in combination with it. It includes more than 150,000 collocations for nearly 9,000 headwords and over 50,000 examples that illustrate collocation context, including, in parts, information on grammar and register.

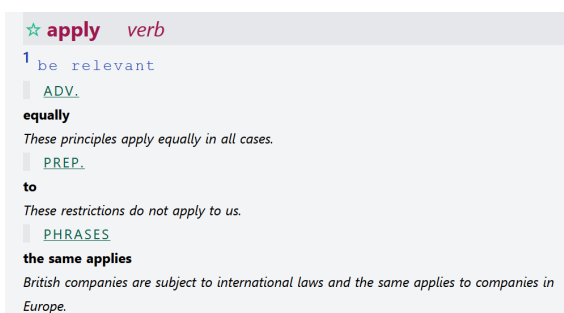


Figure 5: OZDIC: example *apply* (verb).

The lexical entry shown in Figure 5 is divided into several patterns with different associated senses, and this can be made explicit with **OntoLex-Core**:

```
oz:apply-v a ontollex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:verb ;
  ontollex:sense oz:apply-v-sensel ;
  ontollex:canonicalForm [ ontollex:writtenRep "apply"@en ] .
oz:apply-v-sensel skos:definition "be relevant" .
```

²⁴<https://ozdic.com/>

The above statements can be further enriched with morphosyntactic information about the collocation and its parts:

```
oz:equally-adv a ontolex:LexicalEntry;  
  lexinfo:partOfSpeech lexinfo:adverb ;  
  ontolex:canonicalForm [ ontolex:writtenRep "equally"@en ] .
```

As standard lexical resources for English treat `:apply-v` as a lexical entry, and OZDIC does not explicitly distinguish MWEs, phrasal expressions, and syntactic patterns, we model *apply-equally* as a FrAC collocation, assuming that this reflects corpus evidence. With FrAC, attestations (and, subsequently, collocation scores) can also be provided.

```
oz:apply-equally a frac:Collocation, rdfs:Seq ;  
  rdf:_1 oz:apply-v-sense1; rdf:_2 oz:equally-adv ;  
  frac:attestation [  
    frac:quotation "These principles apply equally in all cases." ;  
    frac:corpus <http://www.natcorp.ox.ac.uk/> ] ;  
  frac:head :apply-v-sense1 .
```

Note that here we include the information (given as a statement on the OZDIC website) that the collocations in the dictionary are grounded in the British National Corpus by making use of `frac:attestation` (for corpus evidence);²⁵ the alternative, in cases of examples constructed without provenance, is to use `lexicog:usageExample`. Although OZDIC provides no other corpus-based information at this point in time, this is a sufficient criterion to recommend modelling with FrAC.

Without that statement or the need to encode the source of collocations, an alternative modelling with **decomp** seems feasible:

```
:apply-equally a decomp:Component;  
  decomp:constituent :apply-v , :equally-v ;  
  rdf:_1 :apply-v ; rdf:_2 :equally-adv .
```

Note, however, that this modelling is deficient in that we cannot directly refer to `:apply-v-sense1`, but only to its lexical entry. At the same time, `lexicog:usageExample` cannot be used because the domain of this property is `ontolex:LexicalSense` and not `decomp:Component` (whereas using `frac:attestation` does not have this restriction). So, given the lack of other OntoLex modules

²⁵It is important to note that in FrAC, “corpus evidence” is understood broadly, i.e. is not limited only to linguistic corpora. Since the module has not been published yet and this is one of the issues currently being debated, we recommend referring to the FrAC model specification for the details on what constitutes a `frac:Attestation`.

to adequately reflect the structure of this dictionary entry, we recommend the use of FrAC in this case.

4.3 Enrichment with collocation scores

In Section 4.1, we described the creation of an OntoLex-FrAC resource on the basis of the information contained in a lexicographic resource. With lexical resources, collocation dictionaries, and frequency lists available in OntoLex, we can now trivially bring all of these together. For the OZDIC example in Section 4.2, the collocation “apply equally” can be complemented with n -gram statistics from the corresponding bigram `apply_VERB equally_ADV` in Google Books, with frequencies of the corresponding lexemes and a relative frequency `frac:relFreq` calculated based on the frequency of the collocation and the frequency of its head (“apply”) in all possible inflected forms:

```
gb:apply-equally a gb:GB_2008_coll;
  frac:frequency
    [ rdf:value "16747"; a gb:GB_2008_freq ],
    [ rdf:value "13824"; a gb:GB_2008_doc_freq ] ;
  frac:relFreq "0.00567" ; # = 16747/2954990
  frac:head :apply-v .
oz:apply-equally skos:closeMatch gb:apply-equally .
```

Note that as the OZDIC collocations originate from another corpus, we would produce conflicting metadata entries for `frac:corpus` if we directly related it to the collocation information from Google Book. Thus, we opted to create a new, corpus-specific collocation object and link it to OZDIC by means of `skos:closeMatch`. We suggest `skos:exactMatch` if the collocation contains exactly the same elements (just with a specific basis for calculating their scores), `skos:closeMatch`, if it contains equivalent elements (but, e.g., addressing different aspects, e.g., their entry, form or sense), or `rdfs:seeAlso` if no 1:1 mapping can be established. It is important at this point that this modelling decision is fully independent of whether `:apply-equally` is modelled as `ontoLex:MultiWordExpression`, `decomp:Component`, `lexicog:LexicographicComponent`, `frac:Collocation`: All of these are `frac:Observable`.

5 Discussion and outlook

In this chapter we have focused on describing OntoLex and its modules for the benefit of users who wish to use these vocabularies for modelling multiword

expressions and collocations. Correspondingly, our primary goal has been to give such users some general orientation with regards to the full range of modelling options available in OntoLex for describing such linguistic phenomena in terms of their syntactic, semantic, and morphological structure, as well as in relation to relevant corpus data such as attestations, frequency and collocation scores. For reasons of brevity, we have sought to avoid in-depth descriptions of single use cases, choosing instead to focus on those aspects which will be helpful to anyone modelling similar kinds of data. In terms of an actual resource in which these modelling options have been applied in a comparative manner we can cite a dataset of German compounds (bundled with GermaNet, Hamp & Feldweg 1997). In this case two approaches were taken with a view to meeting two different goals:

- In the first case, with the aim of providing a phrasal analysis without morpheme segmentation; Declerck & Lendvai (2016) describe a shallow representation using **decomp**.
- In the second case, with the aim of facilitating the integration of the dataset with other OntoLex datasets for German morphology; Chiarcos et al. (2022b) describe a representation with morpheme-level segmentation and analysis using **morph**.

As demonstrated above, both of these versions of the dataset – or indeed any other OntoLex data – can be integrated with collocation data as provided, for example by Google N-Grams (see above), the Leipzig Wortschatz portal (Goldhahn et al. 2012), SketchEngine corpora and the Sketch Engine API (Kilgarriff et al. 2014), etc. – regardless of whether their modelling originally made use of **morph**, **decomp** or just plain OntoLex-Core lexical entries.

OntoLex modules can thus be used together in combination (indeed they have been developed for that very purpose). Nonetheless in cases where users of OntoLex are uncertain about which module to use (i.e., their data is not obviously biased towards one module or the other), we recommend that they consider the modules in terms of their order of creation and that such users:

1. Begin by attempting to model their data using **OntoLex-Core** only; if this is insufficient, then
2. Try and apply, in addition, the **synsem**, **decomp**, **vartrans** and **lime** modules; if this also turns out to be insufficient, then

3. Consult, the **lexicog** module; if this is once again to be insufficient, then
4. Consult, the **FrAC** and **morph** modules; if this still fails to meet their modelling needs then
5. As a last resort, join the W3C Community Group where they are invited to discuss their problems or proposed solutions. (Alternatively, create an issue in the respective OntoLex GitHub repository.)²⁶

At the same time, it is advisable to minimise the number of vocabularies involved, so if you *already* know that **morph** will meet your primary modelling needs (e.g., because your dataset or task explicitly requires an emphasis on morphological descriptions), there is no need to combine it with elements of **synsem**, **decomp**, **vartrans**, **lime** or **lexicog** (unless recommended as such in the **morph** vocabulary itself). Such situations of conflict should, however, arise very rarely, because existing modules were taken into account when **lexicog**, **morph** and **FrAC** were developed.

Before closing this chapter, it will be necessary to discuss the advantages and disadvantages of modelling MWEs with OntoLex with reference to the requirements we were initially identified (Section 5.1), and in comparison with pre-RDF technologies (Section 5.2). We also argue for the usability of OntoLex representations of MWEs, with Section 5.3 illustrating this in the case of the elementary task of querying, whereas the final section, Section 5.4, discusses prospective applications.

5.1 Modelling MWEs with OntoLex and RDF technology

This chapter began with the proposal to evaluate current multiword expression modelling strategies in OntoLex according to five criteria. These are the facility with which we can: **identify MWEs** (i.e., to classify them as such); **model the structure of MWEs**; **provide MWE confidence scores**; **facilitate the dynamic prediction** of MWEs with web services and automated tools over existing corpora; and **keep the vocabulary extensible and customizable**, i.e., the capacity of providing concrete usage examples, and detailed, resource-specific metadata or analyses about the respective MWEs, if provided by the underlying resource.

As shown in Table 1, none of the single OntoLex modules discussed here fulfil *all* of these criteria by themselves, but it is important to keep in mind that they are meant to be used *in conjunction* with each other, and in many cases, to build

²⁶<https://github.com/ontolex/>

Table 1: Modelling MWEs with OntoLex. “(+)” indicates partial compatibility.

critierion	OntoLex- Lemon (core)	OntoLex- decomp	OntoLex- FrAC	OntoLex- morph	OntoLex (all)
identification	+	> Lemon	(collocation)	> Lemon	+
structure	-	+	(+)	> decomp	+
scores	-	-	+	-	+
dynamic prediction	-	-	(+)	(+)	(+)
extensible	(+)	(+)	(+)	(+)	(+)

on each other. The **OntoLex-Core** provides the vocabulary to identify MWEs as lexical entries, and in a broader sense, FrAC collocations serve a similar purpose for all combinations of co-occurring expressions. The description of the syntactic and semantic structure of MWEs is handled within **decomp**, and `decomp : subterm` is used for this function in **morph**. FrAC allows for the description of nested collocations (i.e., a collocation that contains another collocation, according to the consideration that collocations are themselves observables), and this can be used to represent phrasal structures – but without any assumptions about their syntactic or semantic interpretability. Collocation scores are a core feature of FrAC, and can be applied to all observables defined in other modules.

As for the dynamic prediction and potential utilisation of these vocabularies for the creation of web services, we focus here on data modelling, and strictly speaking, the vocabularies describe data, not its processing. They are, however, grounded in web standards thus facilitating any subsequent uptake by language technology web services; it should also be borne in mind that such real-world applications have been a driving force throughout the development of OntoLex. In fact, one feature that sets OntoLex apart from competing standards is that it is not tied to a particular serialisation, but that any RDF format (and any format for which an RDF wrapper or injection technology has been designed) can be used, be it a native RDF formalism such as Turtle, JSON, XML, CSV, a triple store, a graph database or a relational database management system, and that data from all of these sources can be trivially transformed using off-the-shelf technology. Competing non-RDF models often claim that they are not inherently tied to any particular serialisation either, but most of the technology developed for working with such models is strongly associated with some preferred format.

As for extensibility, this is another aspect inherent to RDF technology. Standard RDF semantics operate under the open world assumption, i.e., information describing a resource is never taken to be complete by default. Accordingly, native RDF databases are schema-free and data can be extended on demand. At the same time, extensibility does not imply creating novel vocabulary elements in established namespaces. So, while users are encouraged to provide custom vocabulary if necessary, they are also encouraged to put these into separate namespaces rather than polluting the common vocabulary. Such custom vocabularies, if sufficiently mature, and in cases where they enjoy a certain uptake amongst a given user base as well as demonstrating patterns of re-use by third parties, represent the seed for future modules – if there is a consensus in the community and among W3C Community Group chairs about their relevance to OntoLex and its application. But even in this case, this will normally not affect previously published vocabularies: in accordance with general W3C practice, these may be updated at some point in the future, but then, under a different namespace that reflects the time and version of the vocabulary.

5.2 Comparison with non-RDF formalisms

In this section, we give a brief summary of how two other models for lexical resources,²⁷ namely the Lexical Markup Framework (LMF) and the Text Encoding Initiative (TEI), deal with multiword expressions. We have chosen these two because of their influence and popularity in the sector. Indeed OntoLex is historically grounded in LMF,²⁸ the original version of which was published in 2008 by the International Standards Organization (ISO) as standard 24613:2008 and intended as a “standardized framework for the construction of computational lexicons”. LMF originally included a dedicated morphology extension with specific provision for MWEs via the **List of Components** class which allowed for the representation of the “aggregative aspect” of a MWE as well as permitting a recursive description of individual MWE components. This version of LMF also featured a multiword expression pattern extension, which was intended for the representation of the “internal” structure of a MWE and in particular for describing variation within MWEs; this was done via a phrase structure grammar. LMF is currently under revision as a multi-part standard (Romary et al. 2019). However, that part of the new LMF standard which deals with morphology has not

²⁷Although it would be better here to speak of *families* of models for lexical resources.

²⁸LMF is specified using the Unified Modelling Language (UML) and is agnostic about serialisations, although the original standard included an XML serialisation and the latest version of the standard has an associated XML serialisation via TEI. TEI is closely coupled with XML.

yet been published although it is under development. At the time of writing we are aware of no plans to include a MWE pattern component in this latest version of the standard.²⁹ Moreover, LMF does not (and did not in its original version) have a direct equivalent to FrAC and thus lacks specific provision for collocation analysis and the identification of lexicalized MWEs as such: something that is within the scope of applications that consume or produce LMF data.

The XML-based TEI guidelines “define and document a markup language for representing the structural, rendition, and conceptual features of texts”.³⁰ In particular, Chapter 9 of the guidelines provides extensive guidance on encoding dictionaries or related lexicographic resources (Text Encoding Initiative 2022).³¹ In doing so – and notwithstanding the fact that TEI is not intended as a linked data based model – the TEI guidelines provide an informative precedent for the description of collocations in computational lexical resources. We can identify at least three ways in which collocations can be represented in TEI.

One way is to make use of the <colloc> element defined as containing “any sequence of words that co-occur with the headword with significant frequency”.³² <colloc> can be contained in the elements <cit> and <nym> as well as the following elements from the dictionary module: <dictScrap>, <entryFree>, <form> and <gramGrp>.³³ In case the element is located in <gramGrp>, the collocation becomes part of the grammatical information of the entry. Secondly, collocations can also be specified using the <gram> element as is seen in the analysis of French *de médire* in Section 9.3.2 of the TEI guidelines. Thirdly, collocations can be described using the usage element <usg> by specifying the @type attribute of the element as “colloc”.

TEI-Lex0 represents a customisation of the original TEI guidelines with the specific aim of establishing “a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources”³⁴ (Tasovac et al. 2020). TEI-Lex0, as clearly demonstrated by Tasovac et al. (2020), offers much more detailed provision for encoding MWEs than the original TEI guidelines. In particular, by using the <entry> element recursively together with the <gramGrp> element (note that <gramGrp> encodes the information that an entry is a MWE

²⁹Note that the previous version of LMF has been withdrawn as a standard; it is for interest therefore for historical reasons only.

³⁰<https://tei-c.org/guidelines/>

³¹<https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>

³²<https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-colloc.html>

³³In order to see the kinds of attributes which can be used with this element please check the site <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-colloc.html>

³⁴<https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

as well as specifying which type of MWE it is), TEI-Lex0 makes it possible to give a consistent representation to the lexical content of dictionary entries with a distinct visual and/or typographical organisation but similar underlying conceptual organisation. TEI Lex0 recommends a single way of encoding collocates, via `<gram type="collocate">`.

The important insights to be drawn from the TEI guidelines are that (a) there is a demand for modelling collocations in the context of dictionaries (hence multiple, incompatible ways to model it, driven by different use cases and requirements), but that (b) at the moment, the support for modelling collocation scores in this context is severely limited. From the options mentioned above only `<colloc>` allows for the specification of collocation scores by adding a `<certainty>` element and *abusing* its `@cert` attribute, which, however, is only used with human-readable labels in the guidelines,³⁵ but with neither numerical scores nor with a systematic means of defining the type of the collocation score.

With respect to the criteria for MWE and collocation support applied above, it seems that TEI is capable of encoding MWEs and their structure, but that it largely fails at collocation scores. Further, it is extensible by means of ODD customizations. As for dynamic prediction of MWEs, this does not seem to exist as a usage scenario for the TEI, as its deficits in capturing collocation scores reflect. Instead, TEI dictionaries seem to focus on modelling static data, only. In comparison to that, we have argued above that OntoLex captures the demand for MWEs in lexical resources beyond static resources, and shown how FrAC provides the necessary vocabulary for collocation analysis and collocation scores. The current chapter show how OntoLex allows for the seamless integration of MWE-relevant information from different sources, and using SPARQL keywords such as FROM, LOAD and SERVICE, we can even consult data sets (FROM, LOAD) and RDF databases (SERVICE) provided by third parties over the web. This aspect of cross-platform federation is what makes RDF technology truly unique.

What remains to be shown is that it is a technology that can be practically useful, and a minimal requirement for that is *queriability*; this is the topic of the next section.

In summary, then the current version of LMF is limited in its provision for modelling MWEs. It is, however, still missing a morphology part, which when published should somewhat help to improve the situation (even if details are currently short on the ground). TEI on the other hand offers a lot of flexibility in representing MWEs, which can be done via three different elements, namely, `<colloc>`, `<gram>`, and `<usg>`. Indeed in a sense, it offers too much flexibility: there

³⁵<https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-certainty.html>

are too many ways of doing the exact same task. TEI-Lex0 helps to overcome this redundancy, and adds some more expressiveness. However, as we have discussed the result is still limited in terms of provision for collocation scores and dynamic prediction of MWEs.

5.3 Querying MWEs in OntoLex

For any downstream application of lexical data, queriability is the most elementary requirement for a user. Indeed, a key benefit of modelling lexical resources in OntoLex is that they can be processed by standard RDF tools and Linguistic Linked Open Data (LLOD) technology. For Linguistic Linked Open Data, SPARQL provides the possibility to query across data hosted by different providers (SPARQL federation) and across heterogeneous data, i.e., stored in different kinds of technical backends, be it exposed as plain files (SPARQL LOAD), via a web service (SPARQL SERVICE, e.g., an endpoint) or by means of a wrapper technology created around another kind of data source (e.g., a relational data base, using R2RML technology,³⁶ over XML data with GRDDL³⁷ or over JSON data with JSON-LD³⁸ context definitions).

We demonstrate the viability of our modelling for collocations with the application of SPARQL to the OntoLex collocations described above.³⁹

```
SELECT DISTINCT ?collocation ?member ?order
WHERE {
  ?collocation a frac:Collocation ; ?prop ?member .
  FILTER(?prop=rdfs:member || regex(str(?prop),".*#[0-9]+$"))
  OPTIONAL { ?collocation ?nrel ?member .
    FILTER(regex(str(?nrel),".*#[0-9]+$"))
    BIND(replace(str(?nrel),".*#[0-9]+$","$1") AS ?order )
  } } ORDER BY ?collocation ?order ?member
```

This query analyzes two types of membership queries: (1) via `rdfs:member` (2) via filters (`||`) with members in their sequential order (if defined with `rdf:_1`, `rdf:_2`, ...). In other words, this query captures either unordered membership (using `rdfs:member` property) or ordered membership (by filtering on string representation of `rdf:_1`, `rdf:_2`, etc.properties). Note that with RDFS reasoning enabled

³⁶<https://www.w3.org/TR/r2rml/>

³⁷<https://www.w3.org/TR/grddl/>

³⁸<https://www.w3.org/TR/json-ld/>

³⁹Queries were tested with Apache Jena 4.2.0, using the `arq` command line tool. For prefixes and namespaces see the Appendix to this chapter.

at the query engine, `rdfs:member` would also be inferred from `rdf:_1`, etc. For the OZDIC sample data from above, a query with Apache Jena retrieves the following table:

collocation	member	order
:apply-equally	:apply-v-sense	"1"
:apply-equally	:equally-adv	"2"

Appendix B provides additional queries to illustrate the retrieval of all collocations for a given lexical entry and the aggregation of string labels for MWEs. Admittedly, SPARQL queries with aggregation can be complex and difficult to write, particularly for those without technical background in software development or data management. However, in the context of OntoLex, SPARQL is not intended to be exposed to end users, but rather as a backend technology used by technical professionals familiar with the intricacies of querying large data sets.

Although these queries demonstrate the capabilities of OntoLex to address both modelling and information integration challenges in lexical resources in general and for MWEs and collocation analysis in particular, it is clearly a backend technology. What needs to be done at this point is to complement the capabilities of SPARQL with a more user-friendly technical frontend, where queries are generated rather than typed, very much in analogy to how SQL technologies are ubiquitous in modern web technology but almost never exposed to their users. They can play a role, however, in web services that provide or consume lexical data and collocation scores, and in downstream applications that build upon these web services.

5.4 Prospective applications

Identifying and sharing information about MWEs in lexical resources is supported by OntoLex, but unlike its support for RDF, this is not a unique feature among data standards commonly used in this field. What does seem to be unique at the moment is its built-in support for automated collocation analysis, i.e., the inclusion of collocation scores.

Collocations and collocation analysis have been used successfully in information integration for downstream applications. One such application is recommendation systems. Kompan & Bieliková (2011) include collocations into the preprocessing steps used in text mining to create a news recommendation system. The system relies on collocations extracted from the articles' characteristics, e.g., title, content, topics, etc., to recommend news content to users. Chu & Wang (2018)

build a collocation corpus for academic writing in engineering and science fields, then use it to establish a sentence-wide collocation recommendation and error detection system. After extracting collocations, these are classified to create a corpus which is then used to detect collocation errors.

Another application is in computational lexicography, where the well-known platform Sketch Engine currently dominates the market. Sketch Engine provides an API to search and evaluate corpora for automated lexical analyses (“word sketches”), but this is a proprietary system whose services have been disabled for certain groups of users in the past.⁴⁰ With OntoLex-compliant web services, it now becomes possible to develop an open, distributed and provider-independent ecosystem that makes it easier for users to resort to alternative services and data, but that, at the same time, remains inclusive about benefitting from commercial services and data provided by SketchEngine or commercial dictionary providers – that is, if these implement OntoLex specifications in their web services as well. It can thus be viewed as a tool to democratise the market for lexicography, language resources and NLP tools, and to facilitate interoperability and the flow of services and resources between providers and consumers of lexical data and data analytics on the web, for collocation analysis as well as for lexical data in general.

Acknowledgments

The research described in this paper was conducted in the context of the COST Action CA18209 *Nexus Linguarum. European network for Web-centred linguistic data science*. This chapter partially builds on Chiarcos et al. (2022a,c), and we would like to thank GlobalLex 2022 reviewers and audience for feedback and suggestions. Moreover, the authors would like to thank all OntoLex FrAC and OntoLex morph contributors.

The recent development of OntoLex-Morph and OntoLex-FrAC was partially supported by the H2020 Research and Innovation Action Prêt-à-LLOD (2019–2022, ERC grant agreement no. 825182, for Maxim Ionov) and the Early Career Research Group LiODi. Linked Open Dictionaries (2015–2022, BMBF eHumanities programme, for Christian Chiarcos and Maxim Ionov).

⁴⁰This includes changes of licensing conditions (<https://www.sketchengine.eu/access-after-elexis/>) or political reasons (<https://www.sketchengine.eu/news/no-business-as-usual-with-russia-anymore/>).

Abbreviations

API	application programming interface
CSV	comma-separated values
HTTP	Hypertext Transfer Protocol
LexInfo	data category ontology for OntoLex
LLOD	Linguistic Linked Open Data
LMF	Lexical Markup Framework
LOD	Linked Open Data
JSON	JavaScript Object Notation
JSON-LD	JSON for Linked Data
MWE	multiword expression
NLP	natural language processing
ODD	One Document Does it All, schema language for/in TEI-XML
OntoLex	Ontology-Lexica, W3C Community Group and reference vocabulary developed by them
OntoLex-Core	The core module of OntoLex
(OntoLex-)decomp	OntoLex module for decomposition
(OntoLex-)FrAC	OntoLex module for frequency, attestation and corpus-based information
(OntoLex-)lexicog	OntoLex module for lexicography
(OntoLex-)lime	OntoLex module for lexicon metadata
(OntoLex-)morph	OntoLex module for morphology
(OntoLex-)synsem	OntoLex module for syntax and semantics
(OntoLex-)vartrans	OntoLex module for variation and translation
OWL	Web Ontology Language
RDF	Resource Description Language
RDFS	RDF Schema
SKOS	Simple Knowledge Organization Scheme
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TARQL	Tables for SPARQL
TEI	Text Encoding Initiative
TSV	tab-separated values
Turtle	Terse RDF Triple Language
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XML	Extensible Markup Language

RDF namespace prefixes

dbr:	http://dbpedia.org/resource/
dct:	http://purl.org/dc/terms/
decomp:	http://www.w3.org/ns/lemon/decomp
frac:	http://www.w3.org/ns/lemon/frac
lexicog:	http://www.w3.org/ns/lemon/lexicog
lexinfo:	http://www.lexinfo.net/ontology/3.0/lexinfo
lime:	http://www.w3.org/ns/lemon/lime
morph:	http://www.w3.org/ns/lemon/morph
ontolex:	http://www.w3.org/ns/lemon/ontolex
owl:	http://www.w3.org/2002/07/owl
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns
rdfs:	http://www.w3.org/2000/01/rdf-schema
skos:	http://www.w3.org/2004/02/skos/core
synsem:	http://www.w3.org/ns/lemon/synsem
vartrans:	http://www.w3.org/ns/lemon/vartrans

Appendix A OntoLex-FrAC collocation scores

A number of popular collocation scores have been defined as sub-properties of `frac:cscore` within the **OntoLex-FrAC** module, offering clear and established semantics per case. Nonetheless, if the users need to use different scores that are not already provided, they are encouraged to define their own sub-properties, while if they use only one kind of score by a source, they can simply use `rdf:value` along with a `dct:description` to explain the metric. Below, we introduce the existing `frac:cscore` sub-properties along with their mathematical definition. The notations used for the following definitions are:

- x, y - the (head) of the word and its collocate
- $p(x), p(y)$ the probabilities of word x and y
- $p(\neg x) = 1 - p(x)$
- $p(x, y)$ the probability of the co-occurrence of x and y
- $p(x|y)$ the conditional probability of x given y
- N is the sample size

Definition 6.1 (`frac:relFreq`). Relative frequency measures the extent a specific word y occurs together in the collocation of the head word x :

$$\text{relFreq}_x = \frac{p(x, y)}{p(x)}$$

Note that this metric requires `frac:head` to distinguish between the collocation's composing words.

Definition 6.2 (`frac:pmi`). Pointwise Mutual Information (PMI) indicates the degree to which two words in a collocation appear together more than expected under independence. The assumption is that if the words occur more frequently than by chance, then there must be some kind of semantic relationship between them (Role & Nadif 2011). PMI is defined as the log of the ratio of the observed co-occurrence frequency to the frequency expected under independence:

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

Apart from Pointwise Mutual Information well established variants of PMI are also provided with OntoLex-FrAC.

Definition 6.3 (`frac:pmi2`). PMI^2 is a heuristic variant of the PMI measure that aims to increase the influence of the co-occurrence frequency in the numerator and to avoid the characteristic overestimation effect for low-frequency pairs (Role & Nadif 2011):

$$\text{PMI}^2(x, y) = \log \frac{p(x, y)^2}{p(x)p(y)}$$

Definition 6.4 (`frac:pmi3`). PMI^3 uses a higher exponent in the numerator to boost the association scores of high-frequency pairs even further represent a purely heuristic approach (Role & Nadif 2011):

$$\text{PMI}^3(x, y) = \log \frac{p(x, y)^3}{p(x)p(y)}$$

Definition 6.5 (`frac:generalizedPmi`). The generalized PMI^k is also a heuristic approach that tries to correct the bias of PMI towards low-frequency pairs for a given integer $k \geq 1$ and its definition is given by the formula (Role & Nadif 2011):

$$\text{PMI}^k(x, y) = \log \frac{p(x, y)^k}{p(x)p(y)}$$

The parameter k is used to assign more weight to the joint probability $p(x, y)$ since the product of two marginal probabilities, i.e., $p(x)$ and $p(y)$, in the denominator favors pairs with low-frequency words (Role & Nadif 2011).

Definition 6.6 (`frac:npmi`). The Normalized Pointwise Mutual Information (NPMI) normalizes the PMI score in the range $[-1, +1]$, where -1 means that the words never occur together, 0 means that the words are independent, and $+1$ means that there is a complete co-occurrence (Role & Nadif 2011):

$$\text{NPMI}(x, y) = \frac{\text{PMI}(x, y)}{-\log p(x, y)}$$

Definition 6.7 (`frac:pmiLogFreq`). The PMI log Freq (also known as Saliency) is defined as:⁴¹

$$\text{PMI-logFreq}(x, y) = \text{PMI}(x, y) \cdot \log(Np(x, y) + 1)$$

Definition 6.8 (`frac:dice`). Dice coefficient is a metric used to evaluate the collocation of two words x and y and it ranges between 0.0 and 1.0 , where 1.0 indicates complete co-occurrence (Manning & Schütze 1999):

$$\text{Dice}(x, y) = \frac{2p(x, y)}{p(x) + p(y)}$$

Definition 6.9 (`frac:logDice`). The LogDice is an association measure based on Dice, trying to address the problem is that the values of the Dice score are usually very small numbers (Rychlý 2008):⁴²

$$\text{LogDice}(x, y) = 14 + \log_2 \text{Dice}(x, y) = 14 + \log_2 \frac{2p(x, y)}{p(x) + p(y)}$$

Definition 6.10 (`frac:minSensitivity`). Minimum sensitivity is a measure of dependence between word x and word y and it is computed as the minimum of the relative sensitivity of each word (Pedersen 1998):

$$\text{minSensitivity}(x, y) = \min\left(\frac{p(x, y)}{p(y)}, \frac{p(x, y)}{p(x)}\right)$$

In addition to collocation scores, statistical independence tests are employed as scores. To this end OntoLex-FrAC defines additional sub-properties.

⁴¹<https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf>

⁴²<https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf>

Definition 6.11 (`frac:t_score`). The Student's t test (T-score) finds words whose co-occurrence patterns best distinguish two words (Manning & Schütze 1999):

$$T(x, y) = \frac{p(x, y) - p(x)p(y)}{\sqrt{\frac{p(x, y)}{N}}}$$

Definition 6.12 (`frac:chi2`). Pearson's χ^2 test is an alternative to the Student's t test that does not work under the assumption of that the probabilities of words follow the normal distribution (Manning & Schütze 1999):

$$\chi^2(x, y) = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

The observed values O_{ij} are determined using the contingency table of observed frequencies for two words x and y :

	y	$\neg y$
x	$O_{11} = p(x, y)$	$O_{12} = p(x, \neg y)$
$\neg y$	$O_{21} = p(\neg x, y)$	$O_{22} = p(\neg x, \neg y)$

Definition 6.13 (`frac:likelihoodRatio`). The Log Likelihood Ratio test examines the following two alternative hypothesis for the collocation of x and y : $H_1 : p(x|y) = p(x|\neg y) = p(x)$ and $H_2 : p(x|y) \neq p(x|\neg y)$, where H_1 is a formalization of independence, while H_2 is a formalization of dependence. Given that, the Log Likelihood Ratio test is defined as $\log \lambda = \log(L(H_1)/L(H_2))$, where L is the likelihood of each hypothesis (Manning & Schütze 1999). If the ratio is greater than 1, we should prefer H_1 , otherwise we should prefer H_2 . Given that, the Log Likelihood Ratio test has the advantage it is easier to interpret compared to Pearson's χ^2 test and Student's t test.

Furthermore, popular metrics from association rule mining domain are defined as `frac:c_score` subproperties: Within the domain of computational lexicography and corpus linguistics, an association rule $x \rightarrow y$ corresponds to a collocation in that the existence of word x implies the existence of word y .

Definition 6.14 (`frac:support`). Support measures the probability of a rule to appear in the dataset (Larose & Larose 2014):

$$\text{support}(x \rightarrow y) = p(x, y)$$

Definition 6.15 (`frac:confidence`). Confidence measures the probability of a rule to be true (Larose & Larose 2014):

$$\text{confidence}(x \rightarrow y) = \frac{p(x, y)}{p(x)}$$

Definition 6.16 (`frac:lift`). Lift (also known as the interest of a rule) indicates the degree of how often x and y occur together more than expected if they were statistically independent (Larose & Larose 2014):

$$\text{lift}(x \rightarrow y) = \frac{p(x, y)}{p(x)p(y)}$$

Definition 6.17 (`frac:conviction`). The conviction of a rule is the ratio of the expected probability that x occurs without y if x and y are independent, divided by the observed probability of incorrect predictions (Brin et al. 1997):

$$\text{conviction}(x \rightarrow y) = \frac{p(x)p(\neg y)}{p(x, \neg y)}$$

Appendix B Sample queries

As an addendum to §5.3, we model all collocations for a given lexical entry:

```
SELECT DISTINCT ?form ?pos ?collocation
WHERE {
  ?collocation a frac:Collocation ; ?prop ?observable .
  FILTER(?prop=rdfs:member || regex(str(?prop),".*#[0-9]+$"))
  ?entry (ontolex:sense|ontolex:lexicalForm)? ?observable .
  ?entry ontolex:canonicalForm/ontolex:writtenRep ?form .
  OPTIONAL { ?entry lexinfo:partOfSpeech ?pos }
} ORDER BY ?form ?pos ?collocation
```

The second query generates string representations for collocations. This is a bit less straightforward with OntoLex data because string labels are provided for individual words, not necessarily for multiword expressions as a whole – unless an explicit `ontolex:Form` is provided:

```
SELECT DISTINCT ?collocation ?string
WHERE {
  { SELECT ?collocation (GROUP_CONCAT(?wrep; separator=" ") AS ?string)
    WHERE {
```

6 Multiword expressions, collocations and the OntoLex vocabulary

```
{ SELECT ?collocation ?member ?wrep ?order
  WHERE {
    ?collocation a frac:Collocation ; ?prop ?member .
    FILTER(?prop=rdfs:member || regex(str(?prop),".*#[0-9]+$"))
    ?member
      ((^ontolex:sense)?/ontolex:canonicalForm)?/ontolex:writtenRep
      ?wrep.
    OPTIONAL {
      ?collocation ?nrel ?member .
      FILTER(regex(str(?nrel),".*#[0-9]+$"))
      BIND(replace(str(?nrel),".*#[0-9]+$","$1") AS ?order) }
  } GROUP BY ?collocation ?member ?wrep ?order
  ORDER BY ?collocation ?order ?member
} } GROUP BY ?collocation
} }
```

The challenge in this query is that the ordering information retrieved above is to be used in an aggregation (in embedded SELECT statements):

collocation	string	
=====		
:apply-equally	"apply equally"	

References

- Bechhofer, Sean, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah McGuinness, Peter Patel-Schneijder & Lynn Andrea Stein. 2004. *OWL Web Ontology Language Reference*. Tech. rep. World Wide Web Consortium (W3C). <http://www.w3.org/TR/owl-ref/>.
- Bosque-Gil, Julia & Jorge Gracia. 2019. *The OntoLex Lemon lexicography module (Final community group report)*. Tech. rep. W3C. <https://www.w3.org/2019/09/lexicog/>.
- Brewer, Ebenezer Cobham, Alan Isaacs, David Pickering & Elizabeth A. Martin. 1991. *Brewer's dictionary of 20th-century phrase and fable*. Cassell.
- Brin, Sergey, Rajeev Motwani, Jeffrey D. Ullman & Shalom Tsur. 1997. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham (ed.), *ACM SIGMOD international conference on management of data, May 13–15, 1997, Tucson, Arizona, USA*, 255–264. ACM Press. DOI: 10.1145/253260.253325.

- Chiarcos, Christian, Elena-Simona Apostol, Besim Kabashi & Ciprian-Octavian Truičă. 2022a. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, 4018–4027.
- Chiarcos, Christian, Thierry Declerck & Maxim Ionov. 2021. Embeddings for the lexicon: Modelling and representation. In Luis Espinosa-Anke, Dagmar Gromann, Thierry Declerck, Anna Breit, Jose Camacho-Collados, Mohammad Taher Pilehvar & Artem Revenko (eds.), *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*, 13–19.
- Chiarcos, Christian, Christian Fäth & Maxim Ionov. 2022b. Unifying morphology resources with OntoLex-Morph: A case study in German. In *Proceedings of the 13th international conference on language resources and evaluation (LREC-2022)*. Marseille, France.
- Chiarcos, Christian, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan & Ciprian-Octavian Truičă. 2022c. Modelling collocations in OntoLex-FrAC. In Ilan Kernerman & Simon Krek (eds.), *Proceedings of the Globalex workshop on linked lexicography within the 13th Language Resources and Evaluation conference*, 10–18. Paris: European Language Resources Association (ELRA).
- Chiarcos, Christian, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti & Matteo Pellegrini. 2022d. Computational morphology with OntoLex-Morph. In Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos & Maxim Ionov (eds.), *Proceedings of the 8th workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, 78–86. Marseille: European Language Resources Association. <https://aclanthology.org/2022.ldl-1.0>.
- Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck & John Philip McCrae. 2020. Modelling frequency and attestations for Ontolex-Lemon. In *Proceedings of the 2020 Globalex workshop on linked lexicography*, 1–9.
- Chu, Yen-Lun & Tzone-I Wang. 2018. A sentence-wide collocation recommendation system with error detection for academic writing. In Ting-Ting Wu, Yueh-Min Huang, Rustam Shadiev, Lin Lin & Andreja Istenič Starčič (eds.), *ICITL 2018: Innovative technologies and learning (Lecture Notes in Computer Science 11003)*, 307–316. Springer. DOI: 10.1007/978-3-319-99737-7_33.
- Cimiano, Philipp, Paul Buitelaar, John Philip McCrae & Michael Sintek. 2011. Lex-Info: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics* 9(1). 29–51.

- Declerck, Thierry & Piroska Lendvai. 2016. Towards a formal representation of components of German compounds. In Micha Elsner & Sandra Kuebler (eds.), *Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology*, 104–109. Berlin: ACL. DOI: 10.18653/v1/W16-2017.
- Evert, Stefan. 2005. *The statistics of word cooccurrences word pairs and collocations*. Stuttgart: Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. (Doctoral dissertation).
- Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An international handbook*, vol. 2, 1212–1248. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110213881.2.1212.
- Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, UK*.
- Finkbeiner, Rita & Barbara Schlücker. 2019. Compounds and multi-word expressions in the languages of Europe. In Barbara Schlücker (ed.), *Complex Lexical Units*, 1–44. Berlin, Boston: De Gruyter. DOI: 10.1515/9783110632446-001.
- Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monacchini, Mandy Pet & Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation* 43(1). 57–70.
- Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on Language Resources and Evaluation (lrec'12)*, 759–765. Istanbul: European Language Resources Association (ELRA).
- Hamp, Birgit & Helmut Feldweg. 1997. GermaNet: A lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*. <https://aclanthology.org/W97-0802>.
- Hardie, Andrew. 2012. CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3). 380–409.
- Hüning, Matthias & Barbara Schlücker. 2015. Multi-word expressions. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation: An international handbook of the languages of Europe*, vol. 1, 450–467. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110246254-026.

- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubiček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The sketch engine: Ten years on. *Lexicography* 1(1). 7–36.
- Klimek, Bettina, John Philip McCrae, Julia Bosque-Gil, Maxim Ionov, James K. Tauber & Christian Chiarcos. 2019. Challenges for the representation of morphology in ontology lexicons. In *Proceedings of sixth biennial conference on Electronic Lexicography, (eLex 2019)*.
- Kompan, Michal & Mária Bieliková. 2011. News article classification based on a vector representation including words' collocations. In *Advances in intelligent and soft computing*, 1–8. Berlin Heidelberg: Springer. DOI: 10.1007/978-3-642-23163-6_1.
- Larose, Daniel T. & Chantal D. Larose. 2014. Association Rules. In *Discovering Knowledge in Data*, 247–265. John Wiley & Sons. DOI: 10.1002/9781118874059.ch12.
- Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.
- McCrae, John Philip, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2010. *The lemon cookbook*. Tech. rep. <https://lemon-model.net/lemon-cookbook>.
- McCrae, John Philip, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar & Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of fifth biennial Conference on Electronic Lexicography, eLex 2017*. 19–21.
- McCrae, John Philip, Philipp Cimiano, Paul Buitelaar & Georgeta Bordea. 2016. Representing multiword expressions on the web with the OntoLex-Lemon model. In *PARSEME/ENeL workshop on MWE e-lexicons*.
- Mel'čuk, Igor. 2006. Explanatory combinatorial dictionary. In Giandomenico Sica (ed.), *Open problems in linguistics and lexicography*, 225–355. Monza, Italy: Polimetrica.
- Pedersen, Ted. 1998. Dependent bigram identification. In *Proceedings of the 10th Conference on Innovative Applications of Artificial Intelligence (IAAI 1998)*. Madison, WI: AAAI.
- Role, François & Mohamed Nadif. 2011. Handling the impact of low frequency events on co-occurrence based measures of word similarity: A case study of pointwise mutual information. In *Proceedings of the international conference on Knowledge Discovery and Information Retrieval (KDIR 2011)*, 218–223. Setúbal, Portugal: SciTePress. DOI: 10.5220/0003655102260231.

- Romary, Laurent, Mohamed Khemakhem, Anas Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet & Piotr Banski. 2019. LMF reloaded. <http://arxiv.org/abs/1906.02136>.
- Rychlý, Pavel. 2008. A lexicographer-friendly association score. In *RASLAN 2008*, 6–9. Brno: Masarykova Univerzita. <https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf>.
- Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander F. Gelbukh (ed.), *Proceedings of the third international conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, 1–15. Springer.
- Schlücker, Barbara (ed.). 2019. *Complex lexical units: Compounds and multi-word expressions*. Berlin, Boston: De Gruyter. DOI: 10.1515/9783110632446.
- Suchánek, Marek & Robert Pergl. 2020. Case-study-based review of approaches for transforming UML class diagrams to OWL and vice versa. In *2020 IEEE 22nd Conference on Business Informatics (CBI)*, vol. 1, 270–279.
- Tasovac, Toma, Ana Salgado & Rute Costa. 2020. Encoding polylexical units with TEI Lex-o. *Slovenscina 2.0* 8(2). 28–57.
- Text Encoding Initiative. 2022. *P5: Guidelines for electronic text encoding and interchange, Chap. 9 Dictionaries*. Tech. rep. Version 4.4.0. Last updated on 19th April 2022, revision ff9cc28b0. <https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>.

Chapter 7

MWE-Finder: Querying for multiword expressions in large Dutch text corpora

Jan Odijk^a, Martin Kroon^{a,b}, Sheean Spoel^a, Ben Bonfil^a & Tijmen Baarda^a

^aUtrecht University ^bLeiden University

We present MWE-Finder, an application that enables a user to search for multiword expressions (MWEs) in large Dutch text corpora. Components of many MWEs in Dutch can occur in multiple forms, need not be adjacent, and can occur in multiple orders (such MWEs are called *flexible*). Searching for such flexible MWEs is difficult and cannot be done reliably with most search applications. What is needed is a search engine that takes into account the grammatical configuration of the MWE. MWE-Finder is therefore embedded in GrETEL, a treebank search application for Dutch. A user can enter an example of a MWE in a specific canonical form, after which the system searches for sentences in which the MWE occurs, using queries generated automatically from the canonical form. We will describe in detail how the queries for this MWE are derived from the canonical form. The MWE can also be selected from a list of approximately 10k canonical forms for Dutch MWEs that MWE-Finder offers. We will show that MWE-Finder also offers facilities to find examples with unexpected modifiers or determiners on components of the MWE, and that it will yield statistics on the arguments, modifiers and determiners that occur with the MWE and its components.

1 Introduction

Many multiword expressions (MWEs) are flexible in the sense that their components can have different forms, can occur in different orders, or may not be contiguous, with other words appearing between elements of the MWE. This



makes searching for such MWEs in large text corpora difficult. What is needed is a search system that can take all this flexibility into account.

In this chapter we present such a system, called MWE-Finder. This system is specific for the Dutch language, but many aspects of the design of the system are not specific to Dutch or the specific parser used, as we will describe in Section 5.

We made a system for Dutch because this language exhibits flexibility in a wide range of MWEs. This is especially true for verbal MWEs (including proverbs), but also for certain nominal and adpositional MWEs. Searching for Dutch MWEs is thus an excellent and challenging test case for MWE-Finder. In addition, an excellent parser is available for Dutch, Alpino (van Noord 2006), which is also fully integrated in a treebank query application, GrETEL (Augustinus et al. 2017).

MWE-Finder enables a user to find occurrences of a multiword expression in a large Dutch text corpus. MWE-Finder is intended as a tool for any linguist or lexicographer interested in research into MWEs, in particular *flexible* MWEs.

MWE-Finder can be used to address the task of MWE *identification* in the sense of Constant et al. (2017): by using MWE-Finder a researcher can find occurrences of a given MWE easily and in a more reliable way than with other search applications. This will stimulate research into individual MWEs, their variants and their properties, and their frequencies, thereby facilitating research into MWEs in general. The system also creates a good basis for software to automatically annotate large text corpora for MWEs, which not only may be beneficial for linguistic research but also for a variety of natural language processing tools dealing with MWEs.

MWE-Finder uses the DUTch CAnonicalised Multiword Expressions lexical resource (DUCAME) to suggest MWEs to the user. This is a resource containing more than 10,000 MWEs for the Dutch language in a canonical form.

The organisation of this chapter is as follows. We begin with a brief introduction of the notion multiword expression (Section 2). The DUCAME resource is described in more detail in Section 3. MWE-Finder is presented in Section 4. In Section 5 we discuss the potential for extending MWE-Finder to other languages and other parsers. We will end with conclusions (Section 6) and plans for future work (Section 7).

2 Multiword expressions

A MWE is a word combination with linguistic properties that cannot be predicted from the properties of the individual words or the way they have been combined

by the rules of grammar (Odijk 2013b).¹ A word combination can, for example, have an unpredictable meaning (*de boeken neerleggen*, lit. ‘to put down the books’, meaning ‘to declare oneself bankrupt’), an unpredictable form (e.g. *ter plaatse* ‘on location’, with idiosyncratic use of *ter* and *e*-suffix on the noun), or it can have only limited usage (e.g. *met vriendelijke groet* ‘kind regards’, used as the closing of a letter). In a translation context, it can have an unpredictable translation (*dikke darm* lit. ‘thick intestine’, ‘large intestine’), etc.

Note that it is not always easy to determine whether a combination of words is a MWE, because we do not always know the exact properties of the individual component words or what the grammar rules of a language are exactly. So this may require a substantial amount of research.

Words of a MWE need not always be fixed. This can be illustrated with the Dutch MWE *de boeken neerleggen* ‘to declare oneself bankrupt’. The verb *neerleggen* in (1) can occur in all of its inflectional variants (e.g., past participle in (1a), infinitive in (1b), and past tense singular in (1c) and (1d)), and with the separable particle *neer* attached to it (1a, 1b) or separated (1c, 1d). MWEs do not necessarily consist of words that are adjacent, and the words making up a MWE need not always occur in the same order. This expression allows a canonical order with contiguous elements (as in (1a)), but it also allows other words to intervene between its components (as in (1b)), as well as permutations of its component words (as in (1c)), and combinations of permutations and intervention by other words that are not components of the MWE (as in (1d)):

- (1) a. Saab heeft gisteren *de boeken neergelegd*.
 Saab has yesterday the books down.laid
 ‘Saab declared itself bankrupt yesterday.’
- b. Ik dacht dat Saab gisteren *de boeken wilde neerleggen*.
 I thought that Saab yesterday the books wanted down.lay
 ‘I thought Saab wanted to declare itself bankrupt yesterday.’
- c. Saab *legde de boeken neer*.
 Saab laid the books down
 ‘Saab declared itself bankrupt.’
- d. Saab *legde gisteren de boeken neer*.
 Saab laid yesterday the books down
 ‘Saab declared itself bankrupt yesterday.’

¹For a similar but slightly different definition see Sag et al. (2002).

In addition, certain MWEs allow for (and require) controlled variation in lexical item choice, e.g. in expressions containing bound anaphora, where the possessive pronoun varies depending on the subject, as in (2), exactly as in the English expression *to lose one's temper*.

- (2) a. Ik *verloor mijn* / **jouw geduld*.
I lost my / *your patience
'I lost my temper.'
- b. Jij *verloor *mijn* / *jouw geduld*.
You lost *my / your patience
'You lost your temper.'

Of course, not every MWE allows all of these options, and not all permutations of the components of a MWE are well-formed (e.g. one cannot have **Saab heeft neergelegd boeken de*. lit. 'Saab has downlaid books the.').

In Dutch, even proverbs, which have no variable parts, are flexible, because the finite verb occupies a different position in main clauses (3a) than in subordinate clauses (3b), and adverbial modifiers modifying the whole proverb may split the words of the proverb (3c):

- (3) Flexibility of proverbs in Dutch:
- a. *De appel valt niet ver van de boom*.
the apple falls not far from the tree
'The apple never falls far from the tree.'
- b. Hij zegt dat *de appel niet ver van de boom valt*.
he says that the apple not far from the tree falls
'He says that the apple never falls far from the tree.'
- c. *De appel valt immers niet ver van de boom*.
The apple falls after all not far from the tree
'After all, the apple never falls far from the tree.'

This flexible nature of such MWEs makes it difficult to reliably search for such expressions in text corpora. Standard search engines such as Google do not enable the user to systematically search for different word forms of the same lemma. Search applications for Dutch such as OpenSoNaR (van de Camp et al. 2017, de Does et al. 2017) or Nederlab (Brugman et al. 2016) can do this, but it is difficult to formulate a query allowing different orders and interspersed irrelevant words, and the results of such a query will be unreliable. At best, one will find all

instances but at the same time also many cases where all the component words occur but do not make up a MWE. One should be able to search for flexible MWEs in such a way that their grammatical structure is taken into account. This can be done in a treebank, and MWE-Finder enables searching for MWEs in a treebank.

MWEs can contain multiple content words,² but can also contain only a single content word and one or more function words, and can even consist completely of function words. We will not focus here on some classes of MWEs that consist of a content word and one or more function words, such as verbs with obligatory bound reflexive pronouns, verbs with separable particles, verbs with prepositional complements headed by a specific preposition, and combinations thereof, as illustrated in (4), in which the MWE consists of the verb *trekken*, a separable particle (*op*), an idiosyncratically selected adposition (*aan*) and a reflexive pronoun (*zich*):

- (4) Hij heeft *zich* altijd *aan* zijn vriend *op* kunnen *trekken*.
 he has himself always to his friend up can pull
 ‘He has always received support from his friend.’

Such MWEs are already fully dealt with by the grammar used in MWE-Finder.

3 The DUCAME MWE resource

The DUCAME lexical resource is available³ and consists of a reworked version of the DuELME database (Grégoire 2009, 2010, Odijk 2013a) and a new list of MWEs composed by one of the authors on the basis of publicly available sources, which include Stoett (1923), Onze Taal,⁴ VRT website,⁵ Lassy-Small treebank (van Noord et al. 2013), and own collection. DUCAME contains more than 10,000 unique MWEs (many more than DUELME, which had around 5,000).

DUCAME is unique in that it has all the MWEs in a canonical form as described in more detail below. The MWEs also have annotations on properties of their parts. These annotations are based mostly on native speaker intuitions of the developers and have not been tested against large text corpora. MWE-Finder enables carrying out such tests.

² *Content word* is defined here as a word belonging to any of the syntactic categories noun, verb, adjective, or adverb.

³ <https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP>

⁴ <https://onzetaal.nl/schatkamer/lezen/uitdrukkingen> and <https://onzetaal.nl/zoekresultaten?in=advices&zoek=uitdrukking>

⁵ <https://vrrttaal.net/taaladvies-taalkwestie/vaste-uitdrukkingen>

Traditional dictionaries usually include a MWE by providing an example sentence, but it is very difficult for humans and nearly impossible for software to derive the general properties of the MWE from such an example. What is needed is a canonical form from which the properties of the MWE are easy to derive automatically. In addition, the canonical forms should be a well-formed expression of Dutch and should be parsable by automatic parsers.

For single words the canonical form is called the *lemma*, i.e. a specific form of an inflectional paradigm that is used as headword in traditional dictionaries. One can adopt this usage for the head of MWEs as well, and that works fine for many MWEs. However, it does not always work for a MWE with a verb as its head. In Dutch, the lemma of a verb is identical to the infinitive, but several problems arise when one tries to use the infinitive as the lemma for the head of a verbal MWE: first, no overt subjects can appear with an infinitive, so a MWE with an overt subject and an infinitive is an ill-formed expression:⁶

- (5) a. * *De laatste loodjes het zwaarst wegen.*
the last lead.DIM.PL the heaviest weigh
'The tail end is the most difficult.'
- b. * *De schellen iemand van de ogen vallen.*
the scales someone from the eyes fall
'His eyes are opened.'

Furthermore, though the subject must be absent, it is present implicitly and interpreted as an animate actor. If the subject of a MWE is not animate, using the MWE with an infinitival head as the canonical form gives infelicitous results:

- (6) a. ? *iemand de keel uithangen*
someone the throat outhang
'for something to bore someone'
- b. ? *iemand niet kunnen bommen*
someone not can care
'for someone not to care about something'

In order to avoid these problems and at the same time have a canonical form with an infinitive, the canonical forms in this resource are all finite sentences with a form of the future tense auxiliary verb *zullen* 'will' as its main verb, as in (7). These are all well-formed sentences that can in principle be parsed by a parser.

⁶DIM stands for diminutive, PL for plural.

- (7) a. *De laatste loodjes zullen het zwaarst wegen.*
 the last lead.DIM.PL will the heaviest weigh
 ‘The tail end will be the most difficult.’
- b. *De schellen zullen iemand van de ogen vallen.*
 the scales will someone from the eyes fall
 ‘His eyes will be opened.’
- c. *Iets zal iemand de keel uithangen.*
 something will someone the throat out.hang
 ‘Something will bore someone.’
- d. *Iets zal iemand niet kunnen bommen.*
 something will someone not can care
 ‘Someone will not care about something.’

By default, the canonical forms in DUCAME must be interpreted as allowing for the head of the MWE to be modified by determiners and/or other modifiers; a component of the MWE that is not its head cannot be modified by determiners and/or other modifiers individually unless these are themselves components of the MWE. Similarly, it is assumed that only the head of the MWE can occur in different inflectional forms, while other parts of the MWE cannot. Of course, there are many exceptions to this, and these are indicated in DUCAME by means of annotations. The annotations allowed are given in Table 1.

Table 1: Notational devices for annotating a canonical form. The code + can also be combined with * or ! (in any order).

notation	interpretation
* <i>word</i>	<i>word</i> is modifiable/determinable
+ <i>word</i>	<i>word</i> is inflectable
= <i>word</i>	<i>word</i> must occur in the MWE as given
! <i>word</i>	<i>word</i> is not modifiable/determinable
dd:[<i>word</i>]	<i>word</i> must be a definite determiner
< <i>text</i> >	<i>text</i> is interpreted as a freely replaceable argument
0 <i>word</i>	<i>word</i> is not part of the MWE

Arguments of the MWE that can be freely replaced by arbitrary phrases are represented by the indefinite pronouns *iemand* ‘someone’, *iets* ‘something’, and *ergens* ‘somewhere’, where this is possible. One can also use combinations such

as *iemand|iets* or *iets|iemand*, which are to be interpreted as allowing either but most likely with the first alternative. If such words must occur in the MWE as such (i.e. cannot be freely replaced), they can be preceded by the annotation =, as in (8).

- (8) Iemand zal voor =iets tussen iets zitten.
someone will for something between something sit
'Someone will be a factor in something.'

The system of pronouns in natural languages in general and in Dutch in particular is in many respects somewhat arbitrary. So, *iemand* implies a human argument, whereas *iets* implies a nonhuman argument. A distinction between animate and inanimate nonhuman arguments does not exist in the Dutch pronominal system, nor does one between objects and events. For many phrase types there are no pronouns at all, e.g. for adjectival, adverbial and clausal phrases.⁷ Nevertheless, the use of the existing pronouns is easy and rather natural for humans, and the missing pronouns are covered by a special annotation in which an arbitrary phrase surrounded by angled brackets <...> is interpreted as a freely replaceable argument, as in (9).

- (9) Iemand zal <makkelijk> in de omgang zijn.
someone will easy in the interaction be
'Someone will be <easy>-going, <easy> to deal with.'

Bound pronouns such as reflexives and possessive pronouns are represented by the third person singular forms (*zich*, *zichzelf*, *zijn*). If such forms do not vary, one can precede them by the annotation =, as in the expression *op =zich* lit. 'on REFL', 'in itself'. There is (currently) no convention or annotation to specify the antecedent of such bound anaphors.

It sometimes is necessary to include a word in a canonical form to create a natural utterance even if this word does not belong to the MWE. Such words can be preceded by the code 0. This very often occurs in MWEs that have an indefinite subject, which prefer the presence of *er*, as in (10a), and in MWEs that are or contain negative polarity items and that require the presence of a licensing element such as a negative adverb (*niet* 'not'), determiner (*geen* 'no') or pronoun (e.g. *niemand* 'nobody'), e.g., in the MWE with canonical form *dd:[die] vlieger zal 0niet opgaan*. In (10b) the negative adverb *niet* cannot be absent, but it is arguably not part of the MWE, as shown by (10c) in which the negative pronoun *niemand*

⁷Sometimes it is possible to use pronouns for noun phrases to refer to these but there are no pronouns that can actually replace them.

in the main clause is the licensing element for the negative polarity MWE in the subordinate clause.⁸

- (10) a. 0Er zal iets *op het spel staan*.
 there will something on the game stand
 ‘Something will be at stake.’
- b. Die *vlieger* zal **(niet) opgaan*.
 that kite will not up.go
 ‘That won’t wash.’
- c. Niemand denkt dat die *vlieger opgaat*.
 nobody thinks that that kite up.goes
 ‘Nobody thinks that that will wash.’

4 MWE-Finder

MWE-Finder enables a user to search for occurrences of a MWE in a treebank based on an example MWE in the canonical form as described in Section 3. It is embedded in GrETEL, an existing web application for searching Dutch treebanks (Augustinus et al. 2012, 2017, Odijk et al. 2018). The distinguishing feature of GrETEL is its query-by-example feature. In its regular search mode, it leads the user through a number of steps to get from an example sentence to search results and analysis of the search results:

1. *Example*: A user can enter a natural language example that illustrates the construction they are interested in.
2. *Parse*: The Alpino parser (Bouma et al. 2001, van der Beek et al. 2002) parses the example sentence.
3. *Matrix*: The user indicates which words of this example are crucial for the construction, and how each word should be generalised from. Based on this the parse tree of the example sentence is transformed into an XPath query.
4. *Treebanks*: The user can select one or more treebanks to search in.
5. *Results*: The XPath query is applied to the selected treebank(s) and the results are provided as a list of sentences with matches.

⁸The notation **(...)* means that leaving out the parts between the round brackets yields ill-formedness; the notation *(*...)* means that including the part between the brackets leads to ill-formedness.

6. *Analysis*: The results can be further analysed in a graphical interface to a pivot table for properties of the nodes in the query in combination with any available metadata.

A second important feature of GrETEL is that one can upload one's own text corpus, which is then automatically parsed and made available as a treebank to search in.

MWE-Finder is part of version 5 of the web application GrETEL, available in a first version since the end of 2022.⁹ Thanks to this integration, MWE-Finder has access to all GrETEL features, and supports all treebanks that are included in GrETEL as well as the possibility of uploading one's own text corpora. In Sections 4.1 through 4.4 we describe the user interface and the query generation process of MWE-Finder, as well as a number of changes we had to make in GrETEL's backend. In Section 4.2 we illustrate the use of MWE-Finder by means of a concrete example.

4.1 User interface

MWE-Finder partially mimics the structure of GrETEL's main search functionality. It distinguishes the following steps: *Canonical Form* (cf. GrETEL's *Example step*), *Treebanks*, *Results*, and *Analysis*. It currently lacks the *Parse* step and the *Matrix* step.

MWE-Finder enables the user to enter a MWE example, just like GrETEL, though it must be in the canonical form as described in Section 3. The user thereby implicitly formulates a hypothesis about the properties of this MWE. The annotations on the example specify how the system should generalise from this example, so these annotations can be seen as a different way of implementing the *Matrix* step.

The MWEs contained within DUCAME have been included in a drop-down list and are directly searchable within the MWE-Finder. The user can also enter a new MWE, provided that it complies with the conventions for MWE canonical forms (Figure 1).

As a concrete example, suppose that the user selects the canonical form (11):

- (11) Iemand zal de kat uit de boom kijken.
someone will the cat from the tree watch.
'Someone will wait and see.'

⁹<https://gretel5.hum.uu.nl>

Multiword Expressions

Canonical form > [Treebanks](#) > [Results](#) > [Analysis](#)

Canonical form expressions

Showing 20 out of 32 matching known expressions:

iemand zal Oniet voor de kat zijn

iemand zal als de kat om de hete brij lopen

als de katten muizen dan zullen ze Oniet mauwen

bij nacht zullen alle katten grauw zijn

iemand zal de kat bij het spek zetten

iemand zal de kat de bel aanbinden

iemand zal de kat in de kelder metselen

iemand zal de kat in het donker knijpen

de kat zal om der wille van het smeer de kandeleeer likken

iemand zal de kat op het spek binden

iemand zal de kat uit de boom kijken

iemand zal de kat uit de boom zien

een benauwde kat zal rare sprongen maken

iemand zal een kat in de zak kopen

iemand zal een *+kater c:hebben

het eerste gewin zal kattedespinn zijn

het katje van de baan

Figure 1: The first step is to choose a MWE from the DUCAME list of canonical forms or to provide a new MWE.

After the MWE has been selected or entered, the system automatically generates three queries to search for occurrences of this MWE in a treebank. They correspond to different levels of agreement between the MWE and the sentences of the corpora. These are the *major lemma query*, the *near-miss query*, and the *MWE query*.¹⁰ The query generation process is explained in detail in Section 4.3.

Next, the user can select the treebank or treebanks that the query should be applied to. Once selected, the application switches to the *Results* view where query results are displayed as they arrive from the server. In that view, the user can also switch between the different queries for the chosen MWE or choose to exclude results of finer-grained queries. It is also possible to inspect or manually change the automatically generated XPath queries and retrieve new results (Figures 2 and 3).

In the *Results* view, users can also look at the parse trees for results or toggle extra context (one preceding sentence, one following sentence) to better analyse the occurrences found, just like in GrETEL.

Finally, there is the analysis step, which is identical to the one in GrETEL. For a MWE, one would like to analyse the result set in ways that cannot be achieved by GrETEL's standard analysis component. We are working on a special analysis step for MWEs, in which the system gathers statistics on the components of the MWE, the arguments of the MWE (their grammatical relations and syntactic categories, and their heads), the argument frames¹¹ that occur with the MWE, and about modifiers and determiners for the MWE as a whole and for each of its components. It does this for the results of the MWE query, for the results of the near-miss query, and for the difference between the near-miss query and the MWE query. We have an initial version available but at the time of writing it has not been integrated yet in the actual application.

4.2 Illustration

We illustrate the use of MWE-Finder with a specific example. Suppose we want to investigate the use of the MWE *de dans ontspringen* 'to get off scot-free'. The canonical form as listed in DUCAME (version 1) is in (12):

¹⁰Note that MWE-Finder can identify potential occurrences of a MWE in a treebank. It cannot determine for an expression that is ambiguous between a literal and an idiomatic reading which of these alternative readings is applicable in a specific sentence.

¹¹With *argument frame* we mean a list of (extended relation, syntactic category) pairs for the arguments that the MWE occurs with, where an extended relation is a sequence of grammatical relations. For example, in *Marie brak Piets hart*. lit. 'Marie broke Piet's heart', the argument frame is [(su, NP), (obj1/det, NP)], i.e., it combines with two arguments, a subject NP and a NP functioning as the determiner of the direct object.

Multiword Expressions

[Canonical form](#) > [Treebanks](#) > [Results](#) > [Analysis](#)

Query




Canonical form: **iemand zal de kat uit de boom kijken**

Showing query: 1: Multi-word expression query ▾

1: Multi-word expression query

2: Near-miss query

3: Major lemma query

XPath   

```
//
node[
  node[@rel="obj1" and @cat="np" and count(
    node)=3 and
  node[@rel="det" and @cat="detp" and count(
    node)=1 and
  node[@lemma="de" and @rel="hd" and @pt="lid" and @lwtype="bep"]]
  node[@lemma="kat" and @rel="hd" and @pt="n" and @ntype="soort" and (
  node[@rel="mod" and @cat="pp" and count(
    node)=2 and
```

Figure 2: After selecting the treebanks to search in, the results come in and the user can switch between the three queries that are created based on the selected MWE.

Results: 12

Previous [download icon] [print icon] [share icon] [refresh icon] Next

#	ID	Component	Sentence
1	ep-02-05-14.data.dz:334	Year 2002	Wij moeten de kat uit de boom kijken en zien hoe de biotechnologieën zich ontwikkelen en op welke manier zij ingrijpen in de natuur van de mens .
2	ep-02-10-23.data.dz:1784	Year 2002	De leiders van Europa , Tony Blair en de Deense regering uitgezonderd , geloven dat alles zichzelf van binnenuit zal oplossen , als de diplomaten maar genoeg praatjes verkopen , als we de kat uit de boom kijken en kritiek op de VS uiten , in de hoop dat de terroristen niet toeslaan in een grote Europese stad .
3	ep-05-06-23.data.dz:375	Year 2005	De Britse premier kan nog de kat uit de boom kijken , maar de voorzitter van de Raad kan daarmee niet volstaan .
4	ep-06-05-31.data.dz:888	Year 2006	Er zijn geen nationale debatten gevoerd ; er was geen steun van de Europese instellingen , zeker niet van de Raad , die na de negatieve uitslagen van de referenda in Frankrijk en Nederland het proces stopzette en de kat uit de boom wilde kijken .

Figure 3: A sample of the results for the MWE *Iemand zal de kat uit de boom kijken*. ‘Someone will wait and see.’ for the Europarl corpus (Koehn 2005), part of LASSY Groot (van Noord 2008).

- (12) *Iemand zal de dans ontspringen.*
 someone will the dance escape
 ‘Someone will get off scot-free.’

This canonical form is parsed by the parser in MWE-Finder, resulting in the syntactic structure in Figure 4. In this figure, we omit most attribute value pairs on each node, because there are too many to represent.

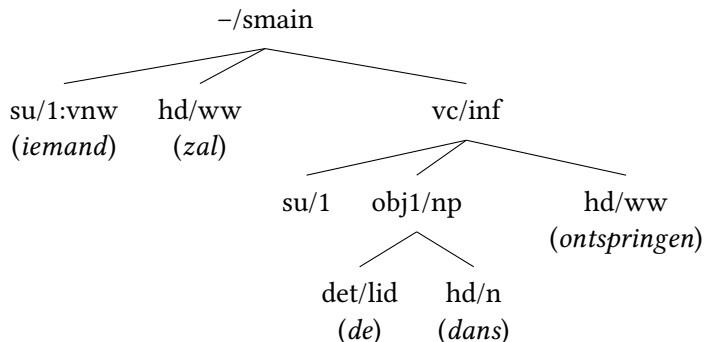


Figure 4: Syntactic structure of *Iemand zal de dans ontspringen.*

The query generation process, described in detail in Section 4.3, first converts this syntactic structure into one or more reduced syntactic structures for the MWE. For this example, there is just one such structure (see Figure 5).

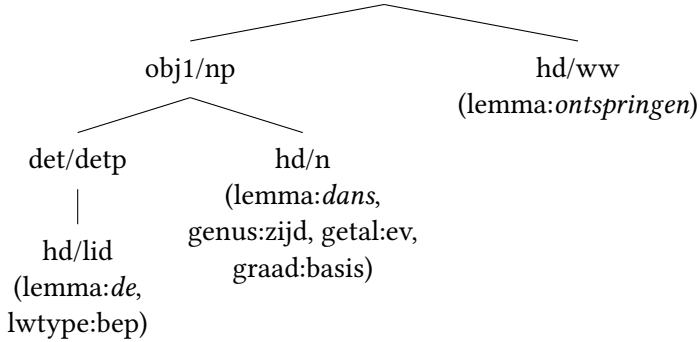


Figure 5: MWE structure of *Iemand zal de dans ontspringen*.

From this the MWE query is generated, shown in Figure 6.

```

//node[
  node[@rel="obj1" and @cat="np" and count(node)=2 and
    node[@rel="det" and @cat="detp" and count(node)=1 and
      node[@lemma="de" and @rel="hd" and @pt="lid" and
        @lwtype="bep"]
    ] and
    node[@lemma="dans" and @rel="hd" and @pt="n" and
      @ntype="soort" and (@genus="zijd" or @getal="mv") and
      @getal="ev" and @graad="basis"]
  ] and
  node[@lemma="ontspringen" and @rel="hd" and @pt="ww"]
]
  
```

Figure 6: The MWE query for *de dans ontspringen*.

When we apply this query to the Mediargus treebank,¹² MWE-Finder finds 1158 hits in over 103 million sentences.

The near-miss query is given in Figure 7. It finds 1271 hits in the Mediargus treebank.

¹²A large treebank with Flemish newspaper text created by Kris Heylen from KU Leuven in 2009.

```
//node[
  node[@rel="obj1" and @cat="np" and
    node[@lemma="dans" and @rel="hd" and @pt="n" and
      @ntype="soort" and (@genus="zijd" or @getal="mv")]
    ] and
  node[@lemma="ontspringen" and @rel="hd" and @pt="ww"]
]
```

Figure 7: The near-miss query for *de dans ontspringen*.

If we exclude the results of the MWE query, which is an option offered by MWE-Finder, we quickly see in the 131 remaining hits that *de dans ontspringen* occurs in variants not predicted by the canonical form that we started with. We list some examples of phrases that the word *dans* occurs with:

different determiners: None,
die ‘that’,
zijn ‘his’;

adjectival modifiers: *gerechtelijke* ‘judicial’,
fiscale ‘fiscal’,
politieke ‘political’;

PP modifiers: *van de bedreigden* ‘of the threatened ones’,
van de sociale verkiezingen ‘of the social elections’.

We also see that the NP headed by *dans* can be the object of two different coordinated verbs, which is possible (we hypothesise) because the verb *ontspringen* is used in its literal meaning in the MWE (i.e., a meaning that it also has outside of this MWE):

- (13) Wie de politieke *dans* gaat leiden of *ontspringen* ...
who the political dance goes lead or escape ...
‘Who will lead or escape the political unpleasant event ...’

All of this clearly suggests that the canonical form that we started with was too strict. We must allow for modification of the MWE component *dans*,¹³ the article

¹³This word appears to have a metaphorical meaning in this MWE, meaning something like ‘unpleasant event’.

de is not a component of the MWE,¹⁴ and ideally we should indicate somehow that the verb *ontspringen* is used in its literal meaning.¹⁵ A better canonical form for this MWE would be *iemand zal Ode *dans ontspringen*, which explicitly allows modification of *dans*, and explicitly states that the determiner *de* is not a component of the MWE. Indeed, the MWE query derived from this canonical form finds 1271 hits, the same number as the near-miss query for the original canonical form. In this way, we can improve upon an initial canonical form mainly based on native speaker intuitions by systematically taking into account corpus data. MWE-Finder makes this possible in a very efficient and user friendly way.

Lastly, the major lemma query (see Figure 8) finds 1309 hits. If we exclude the results of the near-miss query, we have to inspect 38 examples. These are mostly valid instances of the MWE *de dans ontspringen* that have been wrongly parsed by Alpino, but we also find a variant of the MWE, viz. (14) for which we can now add a canonical form to DUCAME.

- (14) Iemand zal *aan de dans ontspringen*.
 someone will to the dance escape
 ‘Someone will get off scot-free.’

In this way, a linguist or lexicographer can easily and efficiently investigate the properties of Dutch MWEs, and improve the description of Dutch MWEs. This process will be even more efficient as soon as the dedicated analysis options have become available.

```
//node[@lemma="dans" and @pt="n"]
/ancestor::alpino_ds/node[@cat="top" and
descendant::node[@lemma="ontspringen" and @pt="ww"]]
```

Figure 8: The major lemma query for *de dans ontspringen*.

4.3 Query generation

Query generation by MWE-Finder involves multiple aspects. In Section 4.3.1 we list and characterise the queries generated. In Section 4.3.2 we describe which

¹⁴Absence of a determiner is generally ill-formed, but this is due to the normal rules of the Dutch grammar, viz. that a singular count noun requires a determiner. This should not be described as a property of the MWE. There are examples in the treebank in which *dans* occurs without a determiner, but these are all examples from headlines which obey a different grammar.

¹⁵A newer version of DUCAME, not described here, has this option.

grammatical properties from the parse of the canonical form end up in the query. Section 4.3.3 lists multiple variants of the MWE structure that must be taken into account. Section 4.3.4 explains how MWE-Finder deals with left-right order. Finally, in Section 4.3.5 we describe the limitations of the approach taken.

4.3.1 Queries

The system processes an input example and interprets it as a canonical form for a MWE: it extracts the annotations and stores them in a data structure, parses the canonical form (with annotations removed) using the Alpino parser, processes any annotations on the canonical form, and then creates three queries: the *major lemma query*, the *near-miss query*, and the *MWE query*. These queries are then applied to a treebank offered by the GrETEL application and selected by the user.

The major lemma query searches for sentences in which at least the lemmas of the so-called major words of the MWE occur (in any grammatical configuration). Major words are the content words if there are at least two in the MWE, and content and function words if there is at most one content word in the MWE. The query yields a superset of the results of both other queries. This query is applied to the full treebank, making use of indexes on the treebank to speed up the process. The major lemma query yields a list of syntactic structures, and can be used to identify the MWE in a grammatical configuration that was not expected at all, to retrieve occurrences of the MWE in sentences that Alpino parsed incorrectly, or to retrieve occurrences of the MWE for which MWE-Finder did not generate the correct other two queries on the basis of the canonical form. The syntactic structures in the output of the major lemma query are adapted in ways described below. The near-miss query and the MWE query are applied to the modified output of the major lemma query.

The near-miss query searches for sentences in which the lemmas of the major words of the MWE occur in the grammatical configuration derived from the canonical form. It can find potential examples of the MWE that deviate from the canonical form provided by showing differences in forms, arguments, modification and determination. It yields a superset of the MWE query results and can be used to fine-tune the hypothesis on the MWE as encoded in the canonical form supplied by the user.

The MWE query finds sentences in which the MWE occurs. This query takes into account the hypothesis on the MWE implied by the canonical form and its annotations supplied by the user.

4.3.2 Grammatical properties

The parse tree for the canonical form contains grammatical properties for each word.¹⁶ These include attributes for the part of speech (*pt*), for the grammatical relation the word has in the structure (*rel*), for the lemma of the word (*lemma*), for the actual form of the word in the utterance (*word*), and for other grammatical properties, among which we distinguish three classes:

Subcategorisation properties: properties to specify a subcategory of the part of speech, e.g. is a pronoun a demonstrative pronoun or a relative pronoun, is an adposition a preposition or a postposition, is a conjunction a coordinate conjunction or a subordinate conjunction, etc.

Interpretable properties: properties that have an influence on the meaning of the utterance, e.g. is a noun singular or plural, what is the mood of the verb, what is the tense of a finite verb, etc.

Purely grammatical properties: e.g. the person and number of a finite verb, the inflectional form of an adjective, the case of a pronoun, etc.

For the inflectable words in a MWE the query will contain a condition on the lemma of the word, its part of speech and any relevant subcategorisation properties. For the uninflectable words it is tempting to formulate the condition in terms of the *word* property, but that would be ill-considered for a variety of reasons. The most important and principled reason has to do with purely grammatical properties such as (structural) *case* or inflectional properties of adjectives. The case of a direct object of a MWE component is not part of the MWE, since the word may occur in different case forms depending on the syntactic configuration, e.g. a phrase may be in nominative case when it has been turned into a subject as a result of passivisation. In MWEs consisting of an adjective and a noun the adjective gets its normal inflectional variants in plural and in definite noun phrases, as illustrated in (15):¹⁷

- (15) a. een *vrolijk*-(*)e *Fransje*
 a gay-E Frans.DIM
 ‘a gay spark’

¹⁶These properties include the so-called D-COI properties (Van Eynde 2005) and various Alpino-specific properties.

¹⁷E stands for the adjectival *e*-suffix. *Frans* is a common Dutch name.

- b. *vrolijk*-(e) *Fransjes*
gay-E Frans.DIM.PL
'gay sparks'
- c. dit *vrolijk*-(e) *Fransje*
this gay-E Frans.DIM
'this gay spark'

A second reason for not formulating the part of the query for uninflectable words in terms of the *word* attribute is that the value of the *word* attribute is how the word actually appears in the text, including capitalisation, missing or extra diacritics, and spelling errors.

Instead, the relevant part of the query is defined in terms of lemma, part of speech, subcategorisation properties and interpretable properties.

4.3.3 Creating modified variants

Creating the query for the canonical MWE is nontrivial, since the algorithm for it must take into account all the conventions for and the annotations on the canonical form provided for the MWE. For reasons described below, modification of the structure is often required. In many cases it is necessary to generate a query that takes into account multiple variants. These variants are required in part due to properties of the Dutch language, in part due to specific properties of the structures that Alpino yields, and in part due to the difficulty of parsing natural language utterances in general. We list a few examples.

4.3.3.1 Single word phrases

For phrases that consist of a single word Alpino yields structures with a node for the word but not for the phrase.¹⁸ This is in accordance with conventions that have been agreed upon in the consortia that have developed treebanks for Dutch (Hoekstra et al. 2003, van Noord et al. 2011), but it is a very unfortunate feature for querying, because it requires a complication or even duplication of most of the queries (Van Eynde et al. 2016: 106–107, Odijk et al. 2017, Odijk 2022); in MWE-Finder this feature is mitigated by expanding the structures of the example MWE and the structures in the major lemma query results to contain a phrasal node above single word phrases (also illustrated in Figure 10).

¹⁸Alpino is, despite what is stated on <https://www.let.rug.nl/vannoord/alp/Alpino/>, not a dependency parser. It is a parser that yields constituent structures with explicitly labelled dependencies. See also Odijk et al. (2017: 283–285).

4.3.3.2 Bare indexed phrases

For words and phrases that play multiple roles in an utterance Alpino yields separate nodes for each role. One of these is a node for the whole phrase (the *antecedent*), whereas the other nodes are nodes with just the property of the grammatical relation and an index attribute (*bare index nodes*). The bare index nodes are coindexed with the antecedent (have the same value for the *index* attribute). This is used for *wh*-movement, control of the subject of an infinitival clause, for subject and object raising, for object to subject movement in passives, and for various kinds of ellipsis. In MWE-Finder bare index nodes are replaced by their antecedent (though their *rel* attribute is retained), both in the major lemma query output structures and in the structure of the example MWE. This is essential for dealing with passivised MWEs where the object has become the subject (see item below), with raising of subject MWE components, as in (16a), and for *wh*-movement of MWE-components, as in (16b):

- (16) a. *De laatste loodjes* zullen *het zwaarst* wegen.
 the last lead.DIM.PL will the heaviest weigh
 ‘The tail end will be the most difficult.’
- b. *Wiens hart* heeft zij *gebroken*?
 whose heart has she broken
 ‘Whose heart did she break?’

The changes made for single word phrases and bare index node expansion are illustrated in Figure 9 (original parse tree) and Figure 10 (parse tree after single word phrase and bare index node expansion).

4.3.3.3 Passivisation

In passivised variants, several changes occur:

- The direct object, if there is one, is turned into a subject;¹⁹
- the subject is left out or turned into a phrase headed by the adposition *door* ‘by’;
- the verb takes on the past participle form;
- a passive auxiliary (*worden* ‘be’ or *zijn* ‘have been’) can be introduced.

¹⁹In Dutch it is sometimes possible to passivise an intransitive verb or a transitive verb without an object, e.g. *er wordt gedanst* ‘there is dancing’, *er wordt gefietst* ‘people are cycling’, *er wordt gebouwd* ‘something is being built/there is construction going on’, prompting a dummy subject *er* ‘there’ (cf. Broekhuis et al. 2020).

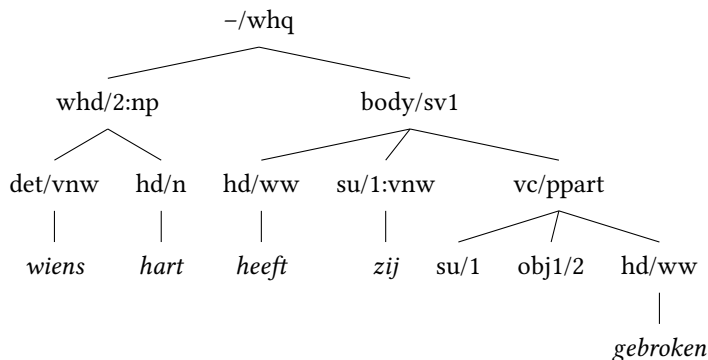


Figure 9: Parse tree of example (16b) *Wiens hart heeft zij gebroken?*. The notation *rel/i:cat* specifies a node with relation *rel*, syntactic category *cat* and index *i*. Not all nodes have an index. Bare index nodes have an index but do not dominate lexical material and have no syntactic category; here *su/1* and *obj1/2*.

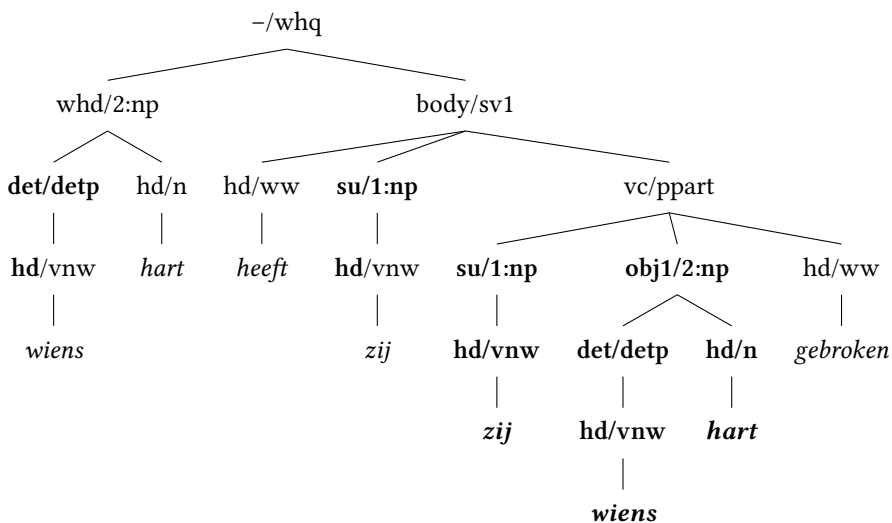


Figure 10: Parse tree of example (16b) *Wiens hart heeft zij gebroken?* after single word phrase expansion and bare index node expansion. The pronouns *wiens* and *zij* are now dominated by a phrasal node. The bare index nodes for the subject and the direct object of the participial phrase have been replaced by their antecedents. Changes are in bold face.

Example (17) illustrates this:

- (17) a. *De boeken* werden door Saab *neergelegd*.
 the books were by Saab down.laid
 ‘Saab declared itself bankrupt.’
- b. Er werd *met de pet naar* *gegoooid*.
 there was with the cap to thrown
 ‘People were mucking around.’

Passive forms of MWEs can be dealt with as follows: free argument subjects are simply not part of the query, since they are not needed at all for identifying a MWE (see also below, under Subjects). Since the object bare index phrase has been replaced by its antecedent, it’s easy to check whether the direct object matches the requirements (in example (17a), whether the direct object equals *de boeken*). Any verb form is accepted by the MWE query, so the past participle also matches. This leaves only the cases where a MWE with a fixed subject can be passivised: in these cases this subject must be replaced by a phrase headed by the adposition *door* in the query. Note that this also accounts for impersonal passives, of which (17b) is an example.

4.3.3.4 Definite pronouns as complements to an adposition

The definite pronouns *het* ‘it’, *dit* ‘this’, and *dat* ‘dat’ as a complement to a preposition are ill-formed or infelicitous. Instead of these, Dutch uses the corresponding R-pronouns (*er*, *hier*, *daar*), with the postpositional variant of the adposition. The R-pronouns precede the adposition and do in fact not have to be adjacent to it:

- (18) a. *Hij *paste een mouw aan het*.
 he fitted a sleeve on it
- b. Hij *paste er een mouw aan*.
 he fitted it a sleeve on
 ‘He found a solution for it.’

For these cases, the query must only allow for the postpositional form of the adposition (e.g. *met* is turned into *mee*); the rest is taken care of by the Alpino grammar itself.

However, if the R-pronoun is adjacent to the adposition, it must be written as one word with the adposition according to the official Dutch spelling rules.²⁰

²⁰https://en.wikipedia.org/wiki/Dutch_orthography.

The simplest way to analyse this is to assume that this is a low-level orthographic convention (grounded in phonological considerations), so that it can be mostly ignored in the syntax (see Rosetta 1994: 115–116 for such an analysis).²¹ But traditional grammar and Alpino deal with these words consisting of an R-pronoun and an adposition (e.g. *eraan* ‘on it’) as independent words with their own part of speech code, so a variant query is generated to cover these cases.

4.3.3.5 Sentential complements to an adposition

The Dutch language does not allow sentential complements to an adposition. This is illustrated in (19a), which has a NP as a complement to the adposition and is well-formed, vs. (19b), which has a subordinate clause as a complement to an adposition and is ill-formed.

- (19) a. *Hij liep tegen veel problemen aan.*
he walked against many problems on
‘He had to face many problems.’
- b. **Hij liep tegen dat hij ziek was aan.*
he walked against that he ill was on
- c. *Hij liep er tegen aan dat hij ziek was.*
he walked it against on that he ill was
‘He had to face the fact that he was ill.’

Instead, the adposition must have the R-pronoun *er* ‘it’ as a complement and changes into a postposition, and the sentential complement is added at the end of the clause, as in (19c). For each query that contains a free argument to an adposition, a variant taking this into account is generated. Also here, an additional variant is generated to cover the cases where *er* and the adposition are written as a single word.

4.3.3.6 Subjects

Subjects can be absent in Dutch utterances in imperative clauses, in cases of topic drop, and in impersonal passives. This is also the case for subjects of MWEs, unless the subject is or contains a fixed component of the MWE. This is illustrated in (20a) for imperatives, in (20b) for topic drop and in (17b) for impersonal passives, repeated here as (20c):

²¹The operation must be syntactic in nature because the R-pronoun and the adposition must only be written as a single word when the R-pronoun is a complement to the adposition.

- (20) a. *Gooi er niet met de pet naar!*
 throw it not with the cap towards
 ‘Don’t muck around!’
- b. *Staat in de sterren geschreven.*
 stands in the stars written
 ‘That is bound to happen.’
- c. *Er werd met de pet naar gegooid.*
 there was with the cap to thrown
 ‘People were mucking around.’

These are accounted for by not including any condition on the subject in the query if it is not and does not contain a fixed component of the MWE. Nonfinite clauses have no overt subject but in most cases they do have a bare index node subject in Alpino structures, so these are not relevant here.

4.3.3.7 Relativisation of MWE-parts

Components of a MWE can sometimes be relativised, especially in the case of support verb constructions, but this is certainly not always the case. Example (21a) contains an example where this is possible (for the MWE (*een*) *poging wagen* ‘to make an attempt’), example (21b) shows an example where this is not possible (for the MWE *de plaat poetsen* ‘to bolt’):²²

- (21) a. *De poging die hij gewaagd had was hopeloos.*
 the attempt that he dared had was hopeless
 ‘The attempt that he had made was hopeless.’
- b. # *De plaat die hij gepoetst had was mooi.*
 the plate that he polished had was beautiful
 ‘The plate that he polished was beautiful.’

MWE-Finder replaces a relative pronoun by its antecedent. As its antecedent it takes the NP that it is contained in with the exclusion of the relative clause²³ but with the addition of an abstract dummy modifier. The antecedent of the relative pronoun *die* ‘that’ is therefore *de dummy poging* in (21a), and *de dummy plaat* in (21b). The presence of the dummy modifier now ensures that relativisation is only allowed when the component of the MWE can be modified (which is the case for *poging* in *een poging wagen*, but not for *plaat* in *de plaat poetsen*).

²²The symbol # means that the idiomatic reading is not possible.

²³This is done to avoid an infinite recursion.

4.3.3.8 NP PP sequences

In expressions in which a noun phrase (NP) is immediately followed by an adpositional phrase (PP) the PP can be a sibling or a child of the NP. Alpino resolves this ambiguity sometimes by selecting the PP as child option, sometimes by selecting the PP as a sibling option. The choice is dependent on several factors, among which the nature of the complement in the PP. In (22) the PP *van iets* is analysed as a child of the NP node dominating *de schuld*, while in (23) the (discontinuous) PP *daar ... van* is a sibling of the NP *de schuld*. We indicated this by means of square brackets in these examples.

- (22) Iemand zal iemand [*de schuld* [*van iets*]] geven.
someone will someone the blame of something give
'Someone will put the blame for something on someone.'
- (23) Iemand zal iemand [*daar*] [*de schuld*] [*van*] geven.
someone will someone there the blame of give
'Someone will put the blame for that on someone.'

For this reason, an alternative structure is generated for nodes headed by a verb that contain an NP which in turn contains a PP. In this alternative structure, the PP is a sibling of the NP.

It is not enough to generate this alternative only for the structure of the MWE that the query is derived from. It must also be applied to the structure of each sentence queried, i.e. for PPs that can be part of the MWE. For example, for the variants *Iemand zal iemand daar van de schuld geven* (with a space between *daar* and *van*) and *Iemand zal iemand van iets de schuld geven*, the PPs *daar van* and *van iets* are analysed as modifiers of the immediately preceding *iemand*, which would lead to a mismatch with the query for the expression *iemand van iets de schuld geven*, as indicated in (24).

- (24) a. Iemand zal [iemand [*daar van*]] *de schuld* geven.
Someone will [someone [there of]] the blame give.
'Someone will put the blame for something on someone.'
- b. Iemand zal [iemand [*van iets*]] *de schuld* geven.
Someone will [someone [of something]] the blame give.
'Someone will put the blame for something on someone.'

4.3.3.9 Adpositional phrases

Adpositional phrases to a verb can get different analyses in Alpino: as a predicative complement, as a locational-directional complement, as an adpositional complement, as a modifier, or as a secondary predicate. The choice is in part dependent on the verb that selects them, but, in the case of ambiguities, also dependent on the disambiguation strategy of Alpino (van Noord 2006), for which it is not easy to predict which selection is made. For PPs dependent on a verb the query is therefore relaxed to accept any of these grammatical relations.

4.3.3.10 Secondary predicates

Modifiers in a clause with a verb cluster are always analysed as modifiers of the deepest embedded verb. However, secondary predicates are always analysed as modifiers of the least embedded verb. If an expression such as (25) with the secondary predicate *als een ketter* is embedded under an auxiliary verb such as *hebben*, as in (26), the phrase *als een ketter* is analysed by Alpino as a modifier to the verb *heeft*, and the MWE will not be found:

- (25) Hij rookt *als een ketter*.
 he smokes like a heretic
 ‘He smokes like a chimney.’
- (26) Hij heeft altijd *gerookt als een ketter*.
 he has always smoked as a heretic
 ‘He has always smoked like a chimney.’

In order to avoid this problem, a special operation is applied to move the secondary predicate to become a modifier of the deepest embedded verb, both in the structures that lead to the query and in the structures of the sentences being queried.

4.3.4 Left-right order

The queries that are generated generally do not check for the left-right order of the components of the MWE, its arguments or modifiers. This is desired since the order of these elements is in most cases not a property of the MWE but follows from the grammar of the language. For this reason MWE-Finder can easily identify the different expressions in (1) as instantiations of the same MWE. Dutch has words that in some cases must be used as a preposition (preceding its complement) and in other cases as a postposition (following its complement), but

even this does not require conditions on order since the distinction is marked by a grammatical feature. Thus, MWE-Finder, without restrictions on left-right order, will correctly not identify (27) as containing the MWE *op de klippen lopen* 'to fail', though it will identify (28) as such:

- (27) Dat zal de klippen op lopen.
that will the cliffs on walk
'That will walk onto the cliffs.' (Not: 'That will fail.')
- (28) Dat zal *op de klippen lopen*.
that will on the cliffs walk
'That will walk on the cliffs.' And: 'That will fail.'

There surely are some MWEs in which the left-right order is a property of the MWE, especially in coordinate structures, e.g. as in (29), but at the time of writing we did not yet implement such restrictions.

- (29) a. dag en nacht
day and night
'during night and day'
- b. # nacht en dag
- c. dames en heren
ladies and gentlemen
'ladies and gentlemen'
- d. # heren en dames

There are also restrictions on left-right order that hold for MWEs but not for literal constructions. For example, *de plaat* in (30) can not be clause-initial under the idiomatic reading though it can be under the literal reading:

- (30) # De plaat heeft hij niet gepoetst.
the plate has he not polished
'He did not polish the plate.' (Not: 'He bolted.')

We did not yet implement such restrictions. We believe that many such restrictions can be dealt with systematically but whether that turns out to be the case still remains to be investigated.

4.3.5 Limitations

MWE-Finder is fully dependent on the syntactic structures generated by the Alpino parser. If Alpino cannot parse a sentence correctly, MWE-Finder will not be able to identify any MWE in it. This is one of the reasons why MWE-Finder includes the major lemma query: this query will find sentences in which the MWE occurs even if Alpino cannot parse it correctly, so a researcher still has data to work with.²⁴ However, this query will also find many sentences in which the MWE does not occur, so it will require more manual work by the researcher. The amount of work is significantly reduced by the option to select the results of a query minus the results of a stricter query, as we showed in Section 4.2. We aim to reduce the amount of manual work required even more by providing statistics on the results and the results minus the results of the other two queries in the dedicated MWE analysis step. In particular, it will provide statistics on the grammatical relation between the lemmas of the major words. However, at the time of writing this has not been integrated in the online version yet.

Alpino may analyse a sentence incorrectly for a wide variety of reasons. One possibility is that the sentence contains a construction that Alpino cannot handle. For example, in the sentence *Hoe goed Afrikaanse muzikanten ook zijn, aan de bak komen ze nauwelijks*. ‘Good though African musicians may be, they hardly get jobs.’²⁵ Alpino can only correctly parse the part *Hoe goed Afrikaanse muzikanten ook zijn*, but it cannot connect it to the rest of the sentence, and as a consequence the MWE *aan de bak komen* ‘to get a job, get a turn’ is incorrectly not identified in this sentence.

MWE-Finder also currently fails to find a MWE if Alpino does not know a word and cannot correctly guess its properties. For example, the word *velen* can be a verb (‘tolerate’) or a pronoun (‘many persons’). As a verb it can only occur in the expression *iets (niet) kunnen velen* ‘not be able to stand something’. Alpino does not know this verb and analyses (31) incorrectly as consisting of a full main clause *hij kan dat niet* ‘he cannot do that’ followed by single word phrase headed by the pronoun *velen* ‘many persons’. Similarly, Alpino does know the verb *smeren*, but only in the sense of ‘to butter’. MWE-Finder can therefore find *smeerde ’m* in (32) when looking for instances of the MWE *’m smeren* ‘to bolt’, because it looks for instances of the verb *smeren* with an object *’m* ‘him’. The problem is that the verb *smeren* ‘to butter’ forms its perfect tense with the auxiliary verb *hebben*, while the verb *smeren* in the MWE *’m smeren* forms its perfect tense with the auxiliary

²⁴Assuming Alpino can at least lemmatise all major words correctly.

²⁵Twente News Corpus (Ordelman et al. 2007), component ad1999, sentence with identifier ad19990108.data.dz:1831 in GrE TEL.

verb *zijn*, as illustrated in (33). The result is that Alpino cannot correctly analyse (33), and MWE-Finder cannot identify it as an occurrence of the MWE *'m smeren*.

(31) Hij *kan* dat niet *velen*.
he can that not stand
'He can't stand it.'

(32) Hij *smeerde* 'm.
he buttered him
'He bolted.'

(33) Hij is 'm *gesmeerd*.
he is him buttered
'He has bolted.'

4.4 Changes under the hood

Under the hood, the backend of GrETEL was largely rewritten to make it more flexible. The existing PHP backend of GrETEL 4 was migrated to Python in combination with the Django framework for web applications,²⁶ which gives us better support for asynchronous tasks and better run-time resource management. This allowed us to improve performance and to better support large corpora and complex queries. The existing Angular frontend of GrETEL 4 was modified to communicate with the new backend and expanded with a new functionality for the MWE-Finder.²⁷

Support for large text corpora is important in the context of MWEs, because word frequencies in natural language have a Zipfian distribution, so that most of the words occurring in the MWEs have very low frequencies. GrETEL 5 still takes a considerable amount of time to search entire corpora, but does so in the background and will cache the counts and results for further usage. We have prepared several existing large corpora for usage in GrETEL, including LASSY Groot (van Noord 2008),²⁸ which includes the 500-million-word SoNaR corpus (Oostdijk et al. 2013), a Wikipedia dump, and the TwNC, a multifaceted Dutch news corpus (Ordelman et al. 2007). GrETEL 5 ships with import scripts for these corpora.

The principles of the existing search mechanism of GrETEL have been retained in GrETEL 5, and they also largely form the basis of how the MWE-Finder is integrated into the application, but with certain deviations. In GrETEL, the corpora

²⁶<https://www.djangoproject.com/>

²⁷<https://angular.io/>

²⁸<https://taalmaterialen.ivdnt.org/download/tstc-lassy-groot-corpus/>

are stored in XML format as they were parsed by the Alpino parser²⁹ in databases of BaseX (Grün 2010),³⁰ a database system for XML documents. GrETEL translates queries created by the user into XQuery/XPath queries that can be executed by BaseX. This search process is relatively slow compared to other search methods, but searching syntactical structures is not possible using common search methods such as simple full text search.

The most important deviation entails that when searching for a MWE, the BaseX databases are always searched using the major lemma query, while the other two queries are executed with the search results of the major lemma query as its basis. The main reason for this is that MWE queries result in complex nested XPath expressions which are not fully optimised by BaseX's query planner.

On the contrary, a major lemma query contains only a handful of content words and makes good use of the indices that BaseX creates for XML attributes. This means that results for a major lemma query can be efficiently retrieved. The results of the major lemma query, which contain all potential matches for the requested MWE, can then be reused for resolving the other more specific queries.

Another reason for searching MWEs based on the major lemma query is that it is necessary to manipulate the Alpino parse trees of the corpora before the other two queries can be run. Those additional manipulations are needed because of the considerations detailed in Section 4.3.3. Such processing steps would be too expensive computationally to run on entire corpora, and are instead run on the result set of the relevant major lemma query. The latter is of a substantially smaller scale. These processing steps are carried out in-memory using the lxml Python library.³¹ The final step is to use the queries to search in the manipulated parse trees, which is done using lxml, as well, thanks to its XPath engine.

Finally, structuring MWE queries around a major lemma query allows query results to be cached, providing a more fluent interactive workflow for users. The user does not see anything of this tiered approach, and instead simply sees the results for the selected MWE and type of query, and can quickly switch between them.

GrETEL is open source and its code is available at GitHub.³² The part of the application that generates queries for MWEs and that performs the tree manipulation is available as a separate Python package, so that it may also be used to create scripts that search treebanks without using GrETEL.³³

²⁹These are in accordance with the alpino_ds DTD.

³⁰https://docs.basex.org/wiki/Main_Page. GrETEL uses BaseX version 9.

³¹<https://lxml.de/>

³²<https://github.com/UUDigitalHumanitieslab/gretel>

³³<https://github.com/UUDigitalHumanitieslab/mwe-query>

5 Other languages

We presented MWE-Finder for the Dutch language, integrated in a specific tree-bank query application (GrETEL) for the Dutch language, which uses a specific grammar and parser for the Dutch language (Alpino). However, it is not difficult to make similar systems for other languages. The minimum requirements to make a variant for a different language are a parser for that language, and a query system that can query the kind of structures that the parser yields. A system for a different language could even be integrated in GrETEL, because GrETEL is in itself not bound to any particular language, as shown by the GrETEL variant for Afrikaans (Augustinus et al. 2016a), and Poly-GrETEL (Augustinus et al. 2016b), which enabled simultaneous querying in multiple languages in a parallel tree-bank.³⁴

Moreover, a MWE-Finder for a different language and a different parser has to have a query generation procedure. The procedure described in §4.3 is to a large extent generic, though of course it has some aspects that are specific to the Dutch language or to the specific parser used. In MWE-Finder, the treatment of single word phrases and the treatment of secondary predicates is entirely idiosyncratic to the parser used. Some aspects are entirely specific to Dutch (definite pronouns and sentential complements to an adposition), though surely each language will have its own peculiarities, even if one would use a cross-language framework for grammatical structures such as the Universal Dependencies framework (Nivre et al. 2016).³⁵ Other aspects will have to be addressed in any grammar/parser but may be implemented in a completely different way in different grammars/parsers. Displacement and control phenomena (with Alpino using bare indexed phrases), passivisation (with Alpino having displaced objects) are concrete examples. But most aspects are completely generic: the treatment of the grammatical properties (§4.3.2), the modified variants (§4.3.3), subjects, relativisation of MWE-parts, NP PP sequences, adpositional phrases, and left-right order are relevant for any language.

In summary, the implementation of MWE-Finder sets an excellent example for the implementation of similar systems for different languages and parsers.

6 Conclusions

We presented the DUCAME resource and the MWE-Finder as useful research instruments for linguistic and lexicological research into MWEs. MWE-Finder

³⁴<http://gretel.ccl.kuleuven.be/poly-gretel/index.php>

³⁵<https://universaldependencies.org/>

makes it possible to reliably and quickly search for occurrences of a MWE despite their flexible nature. The search is based on an example in an annotated canonical form. The system searches not only for the MWE, but also generates and executes two more relaxed queries: the results of the *near-miss query* and especially the difference between the results of the *near-miss query* and the *MWE query* are very useful for evaluating the implicit hypothesis on the nature of the MWE as formulated in the annotated canonical form, and for adjusting it if needed. The *major lemma query*, and especially the difference between the results of this query and the other two enable the user to find occurrences of MWEs that one might not have expected at all, and also acts as a fall back option for cases in which Alpino parses the sentence containing a MWE incorrectly, or if MWE-Finder does not generate the correct other queries from the canonical form.

7 Future work

We aim to finalise the work on the dedicated MWE analysis component and to integrate it in the online application.

We also plan to experiment with a different indexing system than BaseX for the major lemma query. This query searches for a set of lemmas irrespective of their grammatical relation, so it is not necessary to use an index system for this query that can deal with very complex XPath expressions. One of the indexing systems we want to experiment with is Solr/Lucene,³⁶ which has also proven very efficient in OpenSoNaR (de Does et al. 2017) and in Nederlab (Brouwer et al. 2016).

The software behind the system can easily be converted to software to annotate a large corpus for MWEs, and enrich the treebank with metadata on MWE occurrences. We intend to make this software and apply it on a large corpus (e.g., the LASSY-Groot Newspaper corpus; van Noord et al. 2013). We will also write software for converting the metadata on MWE occurrences in the CoNLL-U and Parseme-tsv formats as proposed in the PARSEME consortium.³⁷ This can then form the basis for the manual verification of these annotations and in particular adding missing annotations, and the resulting data may be relevant for a wide range of natural language processing tools dealing with MWEs.

We furthermore aim to extend the annotation system for the canonical forms with special annotations for collocations and support verb constructions, and to extend MWE-Finder so that it can deal with these new annotations.

³⁶<https://lucene.apache.org/> and <https://solr.apache.org/>

³⁷<https://universaldependencies.org/format.html> and <https://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation>

Finally, there is a small number of MWEs that are currently not dealt with correctly with the canonical forms we currently use. We aim to investigate how we can adapt this.

Acknowledgements

The research described in this chapter was carried out in the Datahub SSH project funded by Utrecht University. We thank the anonymous reviewers for their comments, which led to a significant improvement of the original text.

Acronyms and Abbreviations

BaseX	index system for XML documents (index system)
body	relation of the clause in a wh-question (Alpino grammatical relation)
cat	syntactic category (Alpino attribute)
CONLL-U	Computational Natural Language Learning format version U (text corpus format)
det	determiner (Alpino grammatical relation)
detp	determiner phrase (Alpino syntactic category)
DIM	diminutive (grammatical category)
DUCAME	Dutch Canonicalised Multiword Expressions (lexical resource)
DuELME	Dutch Electronic Lexicon of Multiword Expressions (lexical resource)
E	Dutch <i>e</i> -suffix on adjectives (suffix)
GrETEL	Greedy Extraction of Trees for Empirical Linguistics (application)
hd	head (Alpino grammatical relation)
inf	infinitive phrase (Alpino syntactic category)
lid	article (Alpino part of speech code)
lxml	Python module to deal with XML (Python module)
MWE	multiword expression (term)
n	noun (Alpino part of speech code)
np	noun phrase (Alpino syntactic category)
NP	noun phrase (syntactic category)
obj1	direct object (Alpino grammatical relation)
PARSEME	Parsing and Multiword Expressions (project)
PP	adpositional phrase (syntactic category)
ppart	past participle phrase (Alpino syntactic category)
pt	part of speech (Alpino attribute)
rel	grammatical relation (Alpino attribute)

R-pronoun	Dutch pronoun from a particular set, each of which contains an <i>r</i> in it (word class)
smain	main clause (Alpino syntactic category)
su	subject (Alpino grammatical relation)
sv1	Verb-initial clause (Alpino syntactic category)
top	top relation (Alpino grammatical relation)
tsv	tab-separated value file (file format)
TwNC	Twente News Corpus (text corpus)
vc	verbal complement (Alpino grammatical relation)
vnw	pronoun (Alpino part of speech code)
VRT	Vlaamse Radio en Televisie ‘Flemish Radio and Television’ (broadcast organisation in Belgium)
whd	relation of fronted wh-phrase in a question (Alpino grammatical relation)
whq	main wh-question (Alpino syntactic category)
ww	verb (Alpino part of speech code)
XML	eXtensible Mark-up Language (mark-up language)
Xpath	query language for XML-documents (query language)
Xquery	programming language (programming language)

References

- Augustinus, Liesbeth, Peter Dirix, Daniel Van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde & Gerhard Van Huyssteen. 2016a. AfriBooms: An online treebank for Afrikaans. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, 677–682. Portorož, Slovenia: European Language Resources Association (ELRA).
- Augustinus, Liesbeth, Vincent Vandeghinste & Frank Van Eynde. 2012. Example-based treebank querying. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC 2012)*, 3161–3167. Istanbul, Turkey: European Language Resources Association (ELRA).

- Augustinus, Liesbeth, Vincent Vandeghinste, Ineke Schuurman & Frank Van Eynde. 2017. GrETEL: A tool for example-based treebank mining. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the Low Countries*, chap. 22, 269–280. London, UK: Ubiquity. DOI: 10.5334/bbi.22.
- Augustinus, Liesbeth, Vincent Vandeghinste & Tom Vanallemeersch. 2016b. Poly-GrETEL: Cross-lingual example-based querying of syntactic constructions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, 3549–3554. Portorož, Slovenia: European Language Resources Association (ELRA).
- Bouma, Gosse, Gertjan van Noord & Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers* 37(1). 45–59.
- Broekhuis, Hans, Norbert Corver & Riet Vos. 2020. The impersonal passive. In *Taalportaal*. https://taalportaal.org/taalportaal/topic/link/syntax__Dutch__vp__V3_alternations__V3_alternations.3.2.1.2.xml.
- Brouwer, Matthijs, Hennie Brugman & Marc Kemps-Snijders. 2016. A SOLR / Lucene based multi tier annotation search solution. In *Selected papers from the CLARIN annual conference 2016, 26–28 October, Aix-en-Provence*, 29–37. Linköping, Sweden: Linköping University Electronic Press. <https://ep.liu.se/ecp/article.asp?issue=136&article=002&volume=0>.
- Brugman, Hennie, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang & Antal van den Bosch. 2016. Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, 1277–1281. Portorož, Slovenia: European Language Resources Association (ELRA).
- Constant, Mathieu, Gülşen 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. DOI: 10.1162/COLI_a_00302.
- de Does, Jesse, Jan Niestadt & Katrien Depuydt. 2017. Creating research environments with BlackLab. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the Low Countries*, 245–257. London, UK: Ubiquity Press. DOI: 10.5334/bbi.20.
- Grégoire, Nicole. 2009. *Untangling multiword expressions: A study on the representation and variation of Dutch multiword expressions*. Utrecht: Utrecht University. (Doctoral dissertation).

- Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1/2). 23–39. DOI: 10.1007/s10579-009-9094-z.
- Grün, Christian. 2010. *Storing and querying large XML instances*. University of Konstanz. (Doctoral dissertation).
- Hoekstra, Heleen, Michael Moortgat, Bram Renmans, Machteld Schoupe, Ineke Schuurman & Ton van der Wouden. 2003. *CGN Syntactische Annotatie*. CGN report. Utrecht, the Netherlands: Utrecht University. http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/syntax/syn_prot.pdf.
- Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th machine translation summit*, 79–86. Phuket, Thailand.
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Portorož, Slovenia: European Language Resources Association (ELRA).
- Odijk, Jan. 2013a. DUELME: Dutch electronic lexicon of multiword expressions. In Gil Francopoulo (ed.), *LMF: Lexical markup framework*, 133–144. London, UK / Hoboken, US: ISTE / Wiley.
- Odijk, Jan. 2013b. Identification and Lexical Representation of Multiword Expressions. In P. Spyns & J. E. J. M. Odijk (eds.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme* (Theory and Applications of Natural Language Processing), 201–217. Berlin/Heidelberg: Springer.
- Odijk, Jan. 2022. Eenwoordsconstituenten in GrETEL. In *Liber amicorum Francisci Affinii alias Frank Van Eynde*, 143–150. Leuven, Belgium: KU Leuven.
- Odijk, Jan, Martijn van der Klis & Sheean Spoel. 2018. Extensions to the GrETEL Treebank Query Application. In *Proceedings of the 16th international workshop on Treebanks and Linguistic Theories (TLT16)*, 46–55. Prague. <http://aclweb.org/anthology/W/W17/W17-7608.pdf>.
- Odijk, Jan, Gertjan van Noord, Peter Kleiweg & Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the Low Countries*, chap. 23, 281–297. London, UK: Ubiquity Press. DOI: 10.5334/bbi.23.

- Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman. 2013. The construction of a 500 million word reference corpus of contemporary written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch: Results by the STEVIN-programme*, 219–247. Berlin: Springer. DOI: 10.1007/978-3-642-30910-6_13.
- Ordelman, Roeland J.F., Franciska M.G. de Jong, A. J. van Hessen & G. H. W. Hondorp. 2007. TwNC: a multifaceted Dutch news corpus. *ELRA Newsletter* 12(3-4).
- Rosetta, M. T. 1994. *Compositional Translation* (Kluwer International Series in Engineering and Computer Science 273). Dordrecht: Kluwer.
- Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander F. Gelbukh (ed.), *Proceedings of the third international conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, 1–15. Springer.
- Stoett, Frederik August. 1923. *Nederlandsche spreekwoorden, spreekwijzen, uitdrukkingen en gezegden*. 4th edn. Zutphen: W.J. Thieme & Cie. https://www.dbnl.org/tekst/stoe002nede01_01/.
- van de Camp, Matje, Martin Reynaert & Nelleke Oostdijk. 2017. WhiteLab 2.0: A web interface for corpus exploitation. In Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the Low Countries*, 231–243. London, UK: Ubiquity Press. DOI: 10.5334/bbi.19.
- van der Beek, Leonoor, Gosse Bouma & Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde* 7. 353–374.
- Van Eynde, Frank. 2005. *Part Of Speech Tagging En Lemmatisering van het D-COI Corpus*. LASSY Report. Leuven, Belgium: Centrum voor Computerlinguïstiek, KU Leuven. http://www.let.rug.nl/vannoord/Lassy/POS_manual.pdf.
- Van Eynde, Frank, Liesbeth Augustinus & Vincent Vandeghinste. 2016. Number agreement in copular constructions: A treebank-based investigation. *Lingua* 178. 104–126. DOI: 10.1016/j.lingua.2016.02.001.
- van Noord, Gertjan. 2006. At last parsing is now operational. In P. Mertens, C. Fairon, A. Dister & P. Watrin (eds.), *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, 20–42. Leuven, Belgium: ATALA. <https://aclanthology.org/2006.jeptalnrecital-invite.2>.
- van Noord, Gertjan. 2008. Huge parsed corpora in LASSY. In Frank van Eynde, Anette Frank, Koenraad de Smedt & Gertjan van Noord (eds.), *Proceedings of the seventh international workshop on Treebanks and Linguistic Theories (TLT 7)* (LOT Occasional Series 12), 115–126. Groningen: LOT.

- van Noord, Gertjan, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang & Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch* (Theory and Applications of Natural Language Processing), 147–164. Berlin: Springer. DOI: 10.1007/978-3-642-30910-6_9.
- van Noord, Gertjan, Ineke Schuurman & Gosse Bouma. 2011. *Lassy syntactische annotatie (Revision 19455)*. LASSY Report. Groningen. https://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf.

Chapter 8

Collecting and investigating features of compositionality ratings

© Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany

Developing computational models to predict degrees of compositionality for multiword expressions typically goes hand in hand with creating or using reliable lexical resources as gold standards for formative intrinsic evaluation. Not much work however has looked into whether and how much both the gold standards and the computational prediction models vary according to the properties of the compounds within the lexical resources. In the current study, we focus on English and German noun compounds and suggest a novel route to assess the interactions between compound and constituent properties with regard to the compounds' degrees of compositionality. Our contributions are two-fold: (1) a novel collection of compositionality ratings for 1,099 German noun compounds, where we asked the human judges to provide compound and constituent properties (such as paraphrases, meaning contributions, hypernymy relations, and concreteness) before judging the compositionality; and (2) a series of analyses on rating distributions and interactions with compound and constituent properties for our novel collection as well as existing gold standard resources in English and German. Following the analyses we discuss to what extent one should aim for an even distribution of ratings across the pre-specified scale, and to what extent one should take into account properties of the compound and constituent targets when creating a novel resource and when using a resource for evaluation. We suggest as a minimum requirement to balance targets across frequency ranges, and optimally to balance targets across their most salient properties in a post-collection filtering step. Above all, we recommend to assess computational models not only on the full dataset but also with regard to subsets of targets with coherent task-relevant properties.



1 Motivation

Combinations of words are considered multiword expressions (MWEs) in the field of natural language processing (NLP), if they are semantically idiosyncratic to some degree, i.e., the meaning of the combination is not entirely (or even not at all) predictable from the meanings of the constituents (Sag et al. 2002, Baldwin & Kim 2010, Savary et al. 2018). Hence, computational modelling of MWEs has been a long-standing task and is important for both theoretical and applied research, in order to investigate multiword expressions from a large-scale, empirical perspective, and to integrate the compositionality models into NLP applications that require natural language understanding (NLU), such as domain-specific interpretation (Clouet & Daille 2014, HäTTY & Schulte im Walde 2018, HäTTY et al. 2019, Bettinger et al. 2020, HäTTY et al. 2021, Eichel et al. 2023) and machine translation (Carpuat & Diab 2010, Cholakov & Kordoni 2014, Weller et al. 2014, Cap et al. 2015, Salehi et al. 2015b, Gamallo et al. 2019, Dankers et al. 2022).

In the current study, the focus of interest is on noun compounds, such as *climate change* and *crocodile tears* in English, and *Ahornblatt* ‘maple leaf’ and *Fliegenpilz* ‘toadstool’ in German. The representation, processing and modelling of noun compounds has previously received an immense attention across disciplines and languages, e.g., regarding the theoretical definition of compoundhood, typologies of compounds, structural properties of compounds, and compound and constituent meanings (Levi 1978, Plag 2003, Bauer 2017, Schulte im Walde & Smolka 2020, i.a.); regarding the question whether compounds are stored in the mental lexicon and processed as units, via their constituents, or via a dual route (Taft & Forster 1975, Butterworth 1983, i.a.); regarding conceptual combinations of modifiers and heads (Murphy 1990, Wisniewski 1996, Costello & Keane 2000, Benczes 2014, i.a.); regarding the role of compound relations (Gagné 2002, Nastase 2003, Girju et al. 2005, Spalding et al. 2010, i.a.); regarding association and feature norms of noun compounds and their constituents (Roller & Schulte im Walde 2014, Schulte im Walde & Borgwaldt 2015, i.a.); etc.

Standard computational approaches define and compare corpus-based representations of compounds and their constituents, in order to compute the degrees of semantic relatedness as a basis for predicting the degrees of compositionality of the compounds; for example, the representation of a compound such as *climate change* is supposedly more similar to the representations (or a combination of the representations) of the constituents *climate* and *change* than the representation of a more semantically idiosyncratic compound such as *crocodile tears* would be. Such distributional models are rather successful and obtain correlations of $\rho \approx 0.7$ when evaluated against gold standard resources.

Developing computational models of compositionality typically goes hand in hand with creating reliable lexical resources as gold standards for formative intrinsic evaluation. Accordingly, we find datasets of noun compounds with ratings on compositionality across languages, such as English (Reddy et al. 2011b, Cordeiro et al. 2019), German (Schulte im Walde et al. 2013, 2016b), and French and Portuguese (Cordeiro et al. 2019). Not much work however has looked into whether and how much both the gold standards and the prediction models vary according to properties of the targets within the lexical resources. For example, what are the empirical, corpus-based properties of the noun compound targets, such as frequencies and constituent productivities? What are their lexical-semantic properties, such as degrees of ambiguity and concreteness? And how do these properties interact with the compounds' degrees of compositionality? The distributions of target properties and compositionality ratings differ across compound datasets, and potential skewness hinders us from a generalised assessment of prediction models. I.e., does a system's correlation of $\rho \approx 0.7$ hold across targets and target properties, or is this merely an average result and therefore opaque regarding any gold standard subsets? As to our knowledge, up to date only a few computational studies on noun compounds have described the variance of prediction results across compound and constituent properties (Schulte im Walde et al. 2016a, Köper & Schulte im Walde 2017, Alipoor & Schulte im Walde 2020, Miletic & Schulte im Walde 2023), thus pointing out the need for a more systematic investigation.

The current study suggests a novel route to assess the interactions of compound and constituent properties with regard to the compounds' degrees of compositionality, which we consider as indispensable ground knowledge when interpreting the results of computational models. We provide two contributions to move forward both theoretical and computational investigations of compositionality for noun compounds:

- (1) We created a *novel collection of compositionality ratings for 1,099 German noun compounds* where – differently to previous related work – we asked the human judges to provide (a) paraphrases of the compounds' meanings, (b) constituent features contributing to the compounds meanings, (c) judgements on the hypernymy relations between the compounds and their head constituents, and (d) judgements on the concreteness of the compounds and constituents, before they provided their judgements on the compounds' degree of compositionality with regard to the respective constituents. The elaborate information enables us to relate compositionality judgements to a range of compound and constituent properties.

- (2) We present a *series of analyses* on (a) distributions of compositionality ratings, and (b) relations between compositionality ratings and compound and constituent properties (such as frequency, productivity, ambiguity, hypernymy and concreteness). Next to relying on our own novel collection as basis for our study, we also make use of the predominantly used lexical resources of noun compound compositionality for English (Reddy et al. 2011b, Cordeiro et al. 2019) and German (Schulte im Walde et al. 2013, 2016b), and exploit web corpora for the same two languages (Baroni et al. 2009, Schäfer & Bildhauer 2012).

Based on our insights from (1) and (2), we then discuss distributions of compositionality ratings across resources, and to what extent (and how) one should take into account properties of targets when creating a novel resource, and when using a resource in the evaluation of computational models.

In the remainder of this article, Section §2 presents an overview of existing lexical resources with compositionality ratings for noun compounds, as well as standard computational prediction models across languages. In Section §3, our article introduces the creation of the novel gold standard of German compositionality ratings, before we dive into analyses and discussions of rating distributions and rating properties in Section §4.

2 Previous work on compositionality datasets and models

As a starting point for discussing the interactions and potential strategies for optimisations of gold-standard compositionality ratings, we provide an overview of the predominantly used English and German datasets (Section §2.1) and approaches towards predicting degrees of compositionality (Section §2.2).

2.1 Datasets of compositionality ratings

Reddy et al. (2011b) created the probably first dataset with compositionality judgments for noun compounds that were explicitly collected as gold standard ratings to evaluate computational models of compositionality. Henceforth, we will refer to this dataset as REDDY-NN. For the REDDY-NN dataset, Reddy et al. (2011b) selected 90 English noun compounds with two simplex noun constituents. The compound target construction was done such that Reddy et al. distinguished between four classes of modifier and head combinations regarding the constituents'

contributions to the compound meanings,¹ based on heuristics using relations and definitions in WordNet (Fellbaum 1998): a compound was considered compositional with regard to a constituent if it either represented a hyponym of that constituent (e.g., a *swimming pool* is a *pool*), or if the constituent occurred in its definition (e.g., *swimming* occurs in the definition of a *swimming pool*). Then Reddy et al. asked 30 annotators via Amazon Mechanical Turk (AMT) to provide judgements on compositionality ratings for the compound as a whole (which they refer to as “phrase compositionality”), and for the strengths of meaning contributions of the constituents, all on a scale [0, 5] from 0 (clearly non-compositional) to 5 (clearly compositional). The upper part in Table 1 provides a selection of examples from the REDDY-NN target compounds, together with the mean compositionality ratings across the raters and the respective standard deviations. The basic dataset was subsequently extended in various respects: Bell & Schäfer (2013) added semantic relations; Schulte im Walde et al. (2016a) added frequencies and scores for productivity and ambiguity; and Cordeiro et al. (2019), henceforth CORDEIRO-N, extended the dataset to 280 English noun compounds, however varying the modifier word class, and then following the same rating procedure as Reddy et al. (2011b). In our own work, we created two datasets of German noun compounds:

- (1) In Schulte im Walde et al. (2013), we presented a set of 244 German noun-noun compounds with two simplex nominal constituents, based on a larger set of 450 *concrete* noun compounds from von der Heide & Borgwaldt (2009), who had collected compound-constituent compositionality ratings from 30 annotators in a paper-and-pen annotation. We collected and added to the resource between 27–34 compositionality ratings via AMT for the compound as a whole. All ratings were collected on a scale [1, 7] from 0 (clearly non-compositional) to 7 (clearly compositional). Henceforth, we will refer to this dataset as CONCRETE-NN. The lower part of Table 1 provides a selection of examples, together with mean compositionality ratings and standard deviations. The basic dataset was subsequently extended by Schulte im Walde et al. (2016a), who added frequencies and scores for productivity and ambiguity; and Schulte im Walde & Borgwaldt (2015), who compiled and analysed association norms for the concrete compounds and their constituents.

¹As to our knowledge, Libben and his colleagues (Libben et al. 1997, 2003) were the first in psycholinguistics research who systematically categorised noun-noun compounds with nominal modifiers and heads into four groups representing all possible combinations of modifier and head transparency (T) vs. opaqueness (O) within a compound. Examples for these categories were *car-wash* (TT), *strawberry* (OT), *jailbird* (TO), and *hogwash* (OO).

Table 1: Example compounds from REDDY-NN and CONCRETE-NN, with mean compositionality ratings and standard deviations. The compound column refers to compound phrase/whole ratings; the modifier and head columns refer to compound-modifier and compound-head ratings, respectively. Note that the collections use different scales: $[0, 5]$ in REDDY-NN and $[1, 7]$ in CONCRETE-NN.

Compounds	Mean ratings and std. dev.		
	compound	modifier	head
<i>cheat sheet</i>	2.89 ± 1.11	2.30 ± 1.59	4.00 ± 0.83
<i>climate change</i>	4.97 ± 0.18	4.90 ± 0.30	4.83 ± 0.38
<i>couch potato</i>	1.41 ± 1.03	3.27 ± 1.48	0.34 ± 0.66
<i>crocodile tears</i>	1.25 ± 1.09	0.19 ± 0.47	3.79 ± 1.05
<i>diamond wedding</i>	1.70 ± 1.05	0.78 ± 1.29	3.41 ± 1.34
<i>guilt trip</i>	2.19 ± 1.16	4.71 ± 0.59	0.86 ± 0.94
<i>melting pot</i>	0.54 ± 0.63	1.00 ± 1.15	0.48 ± 0.63
<i>night owl</i>	1.93 ± 1.27	4.47 ± 0.88	0.50 ± 0.82
<i>polo shirt</i>	3.37 ± 1.38	1.73 ± 1.41	5.00 ± 0.00
<i>search engine</i>	3.32 ± 1.16	4.62 ± 0.96	2.25 ± 1.70
<i>Ahornblatt</i> ‘maple leaf’	6.03 ± 1.49	5.64 ± 1.63	5.71 ± 1.70
<i>Feuerzeug</i> (lit. ‘fire stuff’) ‘lighter’	4.58 ± 1.75	5.87 ± 1.01	1.90 ± 1.03
<i>Fleischwolf</i> (lit. ‘meat wolf’) ‘meat grinder’	1.70 ± 1.05	6.00 ± 1.44	1.90 ± 1.42
<i>Fliegenpilz</i> (lit. ‘fly mushroom’) ‘fly agaric’	2.00 ± 1.20	1.93 ± 1.28	6.55 ± 0.63
<i>Flohmarkt</i> ‘flea market’	2.31 ± 1.65	1.50 ± 1.22	6.03 ± 1.50
<i>Löwenzahn</i> (lit. ‘lion tooth’) ‘dandelion’	1.66 ± 1.54	2.10 ± 1.84	2.23 ± 1.92
<i>Maulwurf</i> (lit. ‘mouth throw’) ‘mole’	1.58 ± 1.43	2.21 ± 1.68	2.76 ± 2.10
<i>Postbote</i> (lit. ‘mail messenger’) ‘post man’	6.33 ± 0.96	5.87 ± 1.55	5.10 ± 1.99
<i>Seezunge</i> (lit. ‘sea tongue’) ‘sole’	1.85 ± 1.28	3.57 ± 2.42	3.27 ± 2.32
<i>Windlicht</i> (lit. ‘wind light’) ‘lantern’	3.52 ± 2.08	3.07 ± 2.12	4.27 ± 2.36

- (2) In Schulte im Walde et al. (2016b), we presented a dataset of German noun-noun compounds with two simplex nominal constituents. As to our knowledge, this dataset was the first that took properties of the compounds and the constituents into account during the selection of the targets: we induced a balanced set of 180 compounds with low/mid/high modifier productivity and low/mid/high head ambiguity (which we determined as the two most important balancing criteria) from a candidate compound set containing $\approx 150,000$ noun-noun compounds occurring in a large web corpus (Schäfer & Bildhauer 2012). We also created an extended set of 868 compounds by systematically adding all compounds from the original candidate set with either the same modifier or the same head as any of the

compounds in the balanced set. For example, given the compound *Geduldspiel* ‘puzzle’ in the balanced set of compounds we added all compounds from the original candidate set with the modifier *Geduld* ‘patience’, and all compounds with the head *Spiel* ‘game’. We then collected between 8–13 compound–constituent compositionality ratings via AMT, on a scale [0, 6] from 0 (clearly non-compositional) to 6 (clearly compositional). Henceforth, we will refer to the two balanced/unbalanced versions of the dataset containing 180/868 noun–noun compounds as GHOST-NN/S and GHOST-NN/XL, in the same way as in Schulte im Walde et al. (2016a).

Table 2 provides a selection of examples, together with empirical and lexical compound and constituent properties, and mean compositionality ratings. The examples include compounds with the modifiers *Stadt* ‘city’ and *Sonne* ‘sun’ as well as compounds with the heads *Spiel* ‘game’ and *Kette* ‘chain’. The corresponding properties are corpus frequencies for the compounds, modifiers and heads, as well as productivity and ambiguity scores for the constituents, relying on morphological family size (de Jong et al. 2002) and the number of senses defined in GermaNet (Hamp & Feldweg 1997, Kunze 2000), respectively. Semantic relations between modifiers and heads (e.g., in *Machtspiel* ‘power game’, the game is ABOUT power; in *Kartenspiel* ‘card game’, the cards represent the INSTRUMENT in the game) were annotated by the four authors of the paper, adopting the scheme by Ó Séaghdha (2007) using four relations defined by Levi (1978): BE, HAVE, IN, ABOUT; two relations referring to event participants (ACTOR, INST(rument)), and LEX indicating lexicalised compounds.

Overall, the described datasets REDDY-NN and CORDEIRO-N for English as well as CONCRETE-NN and GHOST-NN for German were created on different grounds for target compound selection, i.e., WordNet relations (REDDY-NN and CORDEIRO-N), concreteness (CONCRETE-NN), and partial balancing across empirical and lexical properties (GHOST-NN). The actual collection of human ratings was done similarly across datasets, while varying between paper-and-pen and crowdsourcing as well as the rating scales.

Figure 1 however presents a rather diverse picture regarding the distributions of compositionality ratings across the respective collection ranges. The boxplots show the four quartiles of the rating distributions, with the median lines in the boxes of the interquartile ranges, and the dots referring to outliers. Green boxes refer to compound ratings, blue/red boxes to compound–modifier and compound–head ratings, respectively. For the compound–constituent ratings in

Table 2: Examples of compounds from GHOST-NN/XL with empirical and lexical properties, and mean compositionality ratings.

Compounds	Relation	Frequencies		Productivities		Ambiguities		Ratings		
		compound	modifier	head	modifier	head	modifier	head	modifier	head
<i>Stadthotel</i> 'city hotel'	IN	3,405	4,053,206	1,199,856	543	59	1	1	3.35	5.35
<i>Stadtrand</i> (lit. 'city border') 'suburb'	HAVE	25,099	4,053,206	523,473	543	98	1	2	4.94	4.25
<i>Stadtwerk</i> (lit. 'city plant') 'public services'	ACTOR	107,754	4,053,206	1,354,148	543	366	1	6	3.81	3.69
<i>Sonnenenergie</i> 'solar energy'	INST	25,398	832,636	1,191,333	155	30	3	2	4.58	5.44
<i>Sonnenkönig</i> 'Sun King'	LEX	2,680	832,636	494,221	155	109	3	3	1.94	5.50
<i>Sonnenmasse</i> 'sun mass'	HAVE	3,433	832,636	468,284	155	108	3	3	4.56	4.75
<i>Sonnenscheibe</i> 'solar disc'	BE	3,155	832,636	364,567	155	96	3	4	4.56	3.75
<i>Sonnenseite</i> 'sunny side'	IN	7,279	832,636	5,508,445	155	256	3	6	4.00	4.31
<i>Sonnenstrahl</i> 'sun beam'	HAVE	44,612	832,636	32,182	155	27	3	3	5.13	4.69
<i>Sonnenuhr</i> (lit. 'sun clock') 'sundial'	INST	8,407	832,636	4,507,590	155	63	3	2	3.75	5.31
<i>Kirchspiel</i> (lit. 'church game') 'parish'	LEX	6,583	1,761,187	4,122,168	319	403	3	6	4.44	3.13
<i>Machtspiel</i> 'power game'	ABOUT	4,408	806,162	4,122,168	169	403	2	6	4.63	3.44
<i>Testspiel</i> (lit. 'test game') 'tryout'	BE	37,800	660,169	4,122,168	100	403	3	6	4.25	5.19
<i>Trübspiel</i> (lit. 'mourning game') 'fiasco'	ABOUT	10,763	134,379	4,122,168	77	403	3	6	3.06	2.81
<i>Windspiel</i> (lit. 'wind game') 'wind chimes'	INST	2,284	551,317	4,122,168	88	403	3	6	4.31	2.94
<i>Würfelspiel</i> 'dice game'	INST	4,408	80,371	4,122,168	14	403	2	6	4.94	5.56
<i>Bergkette</i> 'mountain chain'	BE	8,799	564,178	207,479	205	139	2	4	5.13	2.56
<i>Halskette</i> (lit. 'neck chain') 'necklace'	IN	8,707	271,703	207,479	39	139	3	4	3.94	5.44
<i>Handelskette</i> 'trade chain'	INST	6,509	428,611	207,479	240	139	1	4	4.75	3.38
<i>Hotchkette</i> 'hotel chain'	BE	6,410	1,199,856	207,479	134	139	1	4	5.00	3.13
<i>Menschenkette</i> 'human chain'	BE	6,383	8,884,087	207,479	191	139	1	4	4.94	3.75
<i>Produktionskette</i> 'production chain'	HAVE	2,738	579,419	207,479	244	139	2	4	4.69	3.19
<i>Schneekette</i> 'snow chain'	INST	5,167	324,839	207,479	95	139	1	4	4.19	4.21
<i>Zeichenkette</i> (lit. 'character chain') 'string'	BE	8,836	749,903	207,479	62	139	3	4	4.34	2.95

the GHOST-NN variants, 75% of the mean ratings are in the range [4, 6], and the medians are between 4 and 5. CONCRETE-NN is less skewed, but still 75% of all ratings are in the range [3.5, 7]. Only REDDY-NN and the extension CORDEIRO-NN (plots for the latter are in the appendix because they follow similar trends as REDDY-NN) cover a wide range of compositionality ratings. In the next section we will ask whether and how the skewness of the compounds' degrees of compositionality influences the reliability of predictions by computational models.

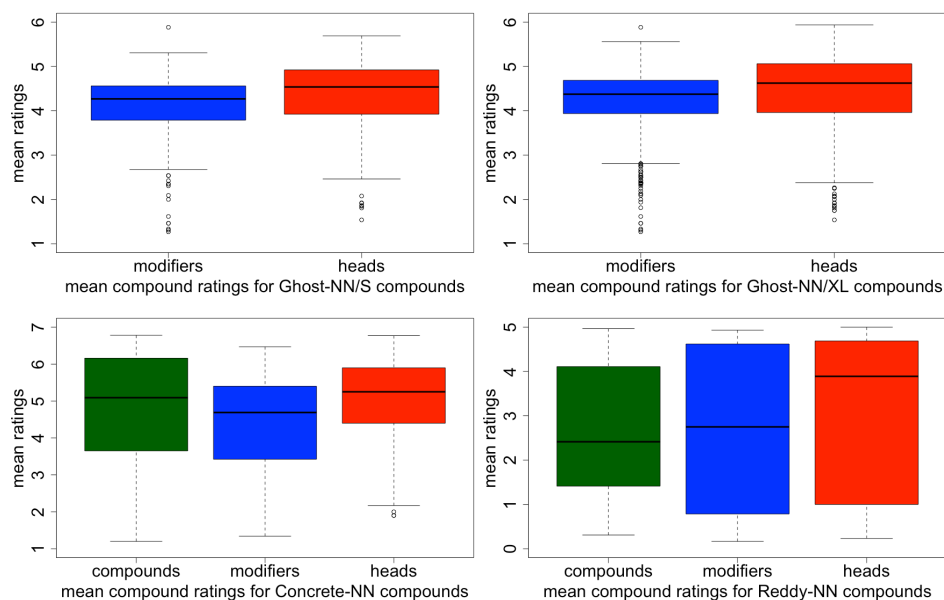


Figure 1: Compositional rating distributions across rating datasets.

2.2 Compositionality prediction models

As introduced above, standard computational approaches define and compare corpus-based representations of compounds and their constituents, in order to determine the degree of semantic relatedness as a basis for predicting the degree of compositionality of the compounds. Existing models generally rely on the distributional hypothesis that the context of a linguistic unit contains indicators for the unit's usage and meaning (Harris 1954, Firth 1957), and thus exploit and represent corpus-based cooccurrences induced from large-scale corpora of the respective language, in combination with mathematical measures of similarity when comparing the representations. The most traditional approaches rely on

distributional count vector spaces, either using window-based or syntax-based cooccurrences (Reddy et al. 2011b,a, Schulte im Walde et al. 2013, 2016a), while later approaches use embeddings as representations (Salehi et al. 2015a, Cordeiro et al. 2019, Alipoor & Schulte im Walde 2020, Miletic & Schulte im Walde 2023). The work by Salehi combined corpus-based textual information with dictionary information (Salehi et al. 2014a) and integrated translation knowledge (Salehi & Cook 2013, Salehi et al. 2014b), and the work in our group extended textual to multimodal approaches (Roller & Schulte im Walde 2013, Köper & Schulte im Walde 2017). While most approaches were directly applied to type-level representations, Bott & Schulte im Walde (2017) applied soft clustering to access the sense level, and Miletic & Schulte im Walde (2023) compared token- and type-level BERT representation layers. The actual predictions of degrees of compositionality then compare the respective representations by computing the cosine distance (or other vector-based distance measures) between vector representations of compounds and vector representations of constituents, or apply composite functions to the vectors of the constituents (such as vector multiplication) before computing the similarity with the compound vector (Mitchell & Lapata 2010, Reddy et al. 2011b, Hermann 2014, Dima et al. 2019, Alipoor & Schulte im Walde 2020, i.a.).

While the exact details of the various approaches are not relevant to the current study, we would like to point out that the majority of approaches predicted the degrees of compositionality across all compound and constituent targets of the respective datasets, i.e., disregarding target subsets and potential influences of such subsets on the prediction. As such, existing compositionality prediction models have overall proven very successful, obtaining Spearman's rank-order correlation coefficients (Siegel & Castellan 1988) of $\rho \approx 0.7$ when evaluated against the gold standard datasets. In the following we present three studies demonstrating that the results differ, however, when compound and constituent properties are taken into account in the evaluation of the models.

Schulte im Walde et al. (2016a) implemented a standard window-based vector space model relying on cooccurrence in a sentence-internal window of ± 20 words, and predicted degrees of compositionality based on the cosine distance measure. For evaluation they used REDDY-NN, CONCRETE-NN and GHOST-NN as well as an English noun compound dataset with semantic relations by Ó Séaghdha (2007). In a preparatory effort, they extended the datasets such that information on compound and constituent frequency, constituent productivity, compound and constituent ambiguity, and semantic relations was available for all English and German resources. Cooccurrences, frequencies and productivities were induced from the respective COW corpora (Schäfer & Bildhauer 2012, Schäfer

2015); ambiguities from WordNet/GermaNet (Fellbaum 1998, Hamp & Feldweg 1997, Kunze 2000), and semantic relations and compositionality ratings were annotated, if not available. Crucially, Schulte im Walde et al. (2016a) then ran their prediction models on all targets within the respective datasets, but also on subsets of targets with extreme properties, such as the least and the most frequent compounds, the least and the most productive constituents, by relation type, etc. Their results showed that – among other insights – the same models make overall better predictions for (i) more frequent compounds, and for (ii) compounds with less frequent, less productive and less ambiguous heads, while (iii) the modifier properties did not have a consistent effect.

In a similar vein, Alipoor & Schulte im Walde (2020) implemented a standard window-based vector space model and word2vec embeddings for English, relying on a sentence-internal window of ± 10 words in the English COW corpus. They focused on the effect of various kinds of dimensionality reductions on compositionality prediction, and they also zoomed into compound subsets regarding compound and constituent properties. Similarly to the study by Schulte im Walde et al. (2016a), they found that in most vector-space variants the predictions (i) were better for mid-/high-frequency compounds in comparison to low-frequency compounds, and (ii) did not behave in a consistent way for modifier properties; but in contrast to the previous work, their predictions were (iii) better for compounds with mid-/high-productivity than low-productivity heads. In addition, they looked into the effect of target compositionality, and found that predictions were (iv) generally better for mid-/high-compositional than low-compositional compound–constituent combinations. Miletic & Schulte im Walde (2023) also zoomed into the influence of frequencies, productivities and ambiguities in our study regarding BERT representation layers. Focusing on head properties of the CORDEIRO-N compounds, we found better compositionality predictions for low-frequency, low-productivity, and low-ambiguity heads across compound and compound–constituent rating predictions.

Finally, Köper & Schulte im Walde (2017) compared multimodal models combining textual and visual vector spaces when predicting degrees of compositionality for German noun compounds and particle verbs. They zoomed into the effects of constituent properties: frequency, ambiguity, concreteness, imageability and compositionality. As in previous work, they did not find consistent effects of modifier properties, but as Schulte im Walde et al. (2016a) they found overall better predictions for (i) compounds with low-frequency and low-ambiguity in comparison to high-frequency/-ambiguity heads; and for (ii) compounds with concrete and imaginable in comparison to abstract and low-imageability heads.

The described studies and their insights clearly demonstrate that – across variants of textual (and also multimodal) vector-space models – compound and con-

stituent properties strongly influence the prediction quality. We are thus asking two questions that we address in the current study. First of all, is there a way to understand better how humans perceive interactions between compound properties and compositionality ratings? We address this question by providing a novel collection (strategy) in Section §3. And secondly, how exactly do compound and constituent properties interact with compositionality ratings, in our novel collection and also in existing datasets? We will address this question by analysing the distributions and correlations of compositionality ratings and compound and constituent property distributions in Section §4.

3 Novel collection: Feature-based compositionality

In this section we present our novel compositionality ratings for German compounds. As target compounds, we rely on the union of targets from the above-described previous German datasets, CONCRETE-NN and GHOST-NN, resulting in a total of 1,099 German noun-noun compounds (i.e., 244 compound targets from CONCRETE-NN and 868 compound targets from GHOST-NN, minus 13 overlapping compound targets). Given that we aimed for a better understanding of what's on an annotator's mind when providing a judgement on a compound's degree of compositionality, we compiled a series of tasks for the annotators to fulfill in addition to providing the actual judgements. In the following we list these tasks, accompanied by the respective motivations. The full annotation guidelines are available in the appendix. The annotators were five graduate students of computational linguistics at the University of Stuttgart.

1. *Compound meaning*: We wanted the annotators to consciously pay attention to the overall meaning of the compound and therefore asked them to paraphrase the compound meaning within a phrase or a sentence. Similar tasks have previously been defined by, e.g., Wisniewski (1996) and Marsh (2015).
2. *Constituent meaning contribution*: Similarly, we wanted the annotators to consciously pay attention to the constituents' meaning components and their contributions to the meaning of the compound. We therefore asked them to explicitly provide one or more features of constituent meaning that contribute to the compound meaning, such as *failure* regarding the contribution of the head *Fehler* 'mistake' to the meaning of the compound *Kunstfehler*.
3. *Super-/sub-ordination (hyponymy/hypernymy)*: We wanted the annotators to be aware of potential hypernymy relationships between compounds and

head constituents, because we hypothesised that a large portion of the compound targets represent sub-ordinate categories (Gagné et al. 2019, 2020). We focused on the compound–head relationship and asked the annotators to judge if the compound is a hyponym (*is a kind*) of the compound head, on a scale [0, 5].

4. *Abstractness/concreteness*: We wanted the annotators to be aware of the concreteness of the compounds and the constituents, because we hypothesised that the degree of concreteness might have an influence on the compositionality of the compound. We therefore asked them to judge about the concreteness (in contrast to abstractness) of compounds and constituents on a scale [0, 5].
5. *Degree of compositionality*: Finally, we wanted the annotators to provide their judgements about the degrees of compositionality of the compounds with regard to their constituents on a scale [0, 5] *after* fulfilling the above-listed tasks about compound and constituent properties.

All annotations are publicly available from <http://www.ims.uni-stuttgart.de/data/feature-comp-nn>, which also includes the spreadsheet for annotation that we gave to the annotators. In the following we provide insights into the various kinds of annotations we collected.

Regarding task 1 (compound meaning), Table 3 shows examples of paraphrases of compound meanings that were provided by the annotators. We can see that the paraphrases are strongly overlapping in some cases, e.g., for the compound *Autozug* ‘car train’ we find four almost identical phrases *Zug, der Autos transportiert* ‘train that transports cars’. Yet, the paraphrases offer different aspects of meanings, such as *schwingen* ‘to swing’, *Instrument* ‘instrument’ and *Dekoration* ‘decoration’ for *Windspiel* (lit. ‘wind game’) ‘wind chimes’. Overall, we judge the paraphrases as useful materials to approach the compound meanings, similarly to dictionary definitions and WordNet glosses.

Regarding task 2 (constituent meaning contribution), Table 4 shows examples of modifier and head features which the annotators considered as contributing to the compound meanings. When comparing these features with the compound paraphrases in Table 3, we can see that the overlap in the materials differs for constituents with more vs. less contributions to compound meaning, e.g., three annotators refer to *Panzer* ‘carapace’ as the meaning contribution of *Schild* ‘shield’ to *Schildkröte* (lit. ‘shield toad’) ‘turtle’, and *Instrument* ‘instrument’ for *Spiel* ‘game’ in *Windspiel* (lit. ‘wind game’) ‘wind chimes’.

Table 3: Examples of compound paraphrases in FEATURE-NN.

<p><i>Autozug</i> ‘car train’</p>	<p><i>(ein)Zug, der Autos transportiert</i> (4 annotators) ‘(a) train that transports cars’ <i>ein Zug für den Fernverkehr, der neben Personen auch Fahrzeuge befördert</i> ‘a train for long-distance traffic that also carries vehicles, next to persons’</p>
<p><i>Eifersucht</i> ‘jealousy’, (lit. ‘eagerness addiction’)</p>	<p><i>Besitzanspruch auf eine Person</i> ‘claim of ownership to a person’ <i>eine Form des Neides im Kontext romantischer Beziehungen</i> ‘a form of jealousy in the context of romantic relations’ <i>Angst die Liebe oder Zuneigung eines Anderen mit jemanden teilen zu müssen</i> ‘fear of having to share someone’s love or affection’ <i>anderer Ausdruck für Neid</i> ‘different expression for jealousy’ <i>Angst jemanden zu verlieren</i> ‘fear to lose someone’</p>
<p><i>Schildkröte</i> ‘turtle’, (lit. shield toad)</p>	<p><i>Reptil mit Panzer</i> ‘reptile with carapace’ <i>eine Reptilienart mit einem charakteristischen Panzer auf dem Rücken</i> ‘a type of reptile with characteristic carapace on back’ <i>Reptilien mit Panzer</i> ‘reptile with carapace’ <i>ein Reptil mit einem harten Panzer um den Torso</i> ‘a reptile with a hard carapace around the torso’ <i>ein Reptil mit einem Panzer</i> ‘a reptile with a carapace’</p>
<p><i>Windspiel</i> ‘wind chimes’, (lit. ‘wind game’)</p>	<p><i>Objekt, das im Wind schwingt</i> ‘object that swings in the wind’ <i>eine Art Instrument, das außerhalb von Gebäuden aufgehängt und vom Wind gespielt wird</i> ‘a kind of instrument that hangs outside buildings and is played by the wind’ <i>Dekoration die im Wind sich bewegt</i> ‘decoration that moves in the wind’ <i>Konstrukt, das sich im Wind bewegt und Geräusche macht</i> ‘construct that moves in the wind and makes sounds’ <i>eine hängende Dekoration, die im Wind Töne erzeugt</i> ‘a hanging decoration that makes sounds in the wind’</p>

8 Collecting and investigating features of compositionality ratings

Table 4: Examples of constituent features contributing to compound meaning in FEATURE-NN.

<i>Bahnhof</i> ‘train station’	<i>Bahn</i> ‘train’	<i>verkehrstechnisch, ziehend</i> ‘transport connecting, pulling’ <i>Bahnverkehr, Zugverkehr</i> ‘rail/train traffic’ <i>Transportmittel</i> ‘means of transport’ <i>Transportmittel, Zug</i> means of transport, train’ <i>Zug</i> ‘train’
<i>Schildkröte</i> ‘turtle’, (lit. ‘shield toad’)	<i>Schild</i> ‘shield’	<i>schildförmig, schützend</i> ‘shield-shaped’, ‘protective’ <i>gepanzert, geschützt</i> ‘armoured’, ‘protected’ <i>mechanischer Schutz</i> ‘mechanical protection’ <i>Panzer, Schutz, robust</i> ‘carapace’, ‘protection’, ‘robust’ <i>gepanzert</i> ‘shielded’
<i>Windspiel</i> ‘wind chimes’, (lit. ‘wind game’)	<i>Wind</i> ‘wind’	<i>windig</i> ‘windy’ <i>Wind</i> ‘wind’ <i>Bewegung in der Luft</i> ‘movement in the air’ <i>Luft, Böe, wehen</i> ‘air’, ‘gust’, ‘blow’ <i>beweglich</i> ‘movable’
<i>Luftzug</i> ‘draught’, (lit. ‘air train’)	<i>Zug</i> ‘train’	<i>ziehend</i> ‘pulling’ <i>bewegt</i> ‘moved’ <i>Transportmittel</i> ‘means of transport’ <i>Bewegung</i> ‘movement’ <i>Richtung</i> ‘direction’
<i>Schildkröte</i> ‘turtle’, (lit. ‘shield toad’)	<i>Kröte</i> ‘toad’	<i>kriechend</i> ‘creeping’ <i>Reptil</i> ‘reptile’ <i>Amphibien die am Wasser leben</i> ‘amphibians that live in the water’ <i>Tier</i> ‘animal’, <i>Frosch</i> ‘frog’ <i>Reptil</i> ‘reptile’
<i>Windspiel</i> ‘wind chimes’, (lit. ‘wind game’)	<i>Spiel</i> ‘game’	<i>spielend</i> ‘playing’ <i>Instrument</i> ‘instrument’, <i>Klang</i> ‘sound’ <i>Vergnügen</i> ‘pleasure’ <i>unterhaltend</i> ‘entertaining’ <i>Musik</i> ‘music’

Regarding task 3 (hypernymy relation between compounds and their head constituents), Table 5 shows examples of mean hypernymy ratings for a subset of the target compounds with heads *Spiel* ‘game’, *Werk* ‘work’; ‘factory’ and *Zug* ‘train’; ‘draught’. The dataset FEATURE-NN contains a total of 39/76/28 compound types (i.e., 39/76/28 different modifiers) with heads *Spiel*, *Werk* and *Zug*, respectively. We can see that these heads strongly differ regarding their hypernymy relation strengths to the respective compounds. Figure 2 shows the distributions of the ratings across all compound heads (red box), and also for only the compounds with the three example heads (orange boxes). The boxplots show that (i) overall we have a target set of compounds that is highly skewed towards super-/subordination, but also that (ii) the hypernymy strength distribution varies according to specific compound heads.

Regarding task 4 (abstractness/concreteness), Figure 3 shows the distributions of the ratings across all compounds, all modifiers and all heads (green, blue and red boxes, respectively, as in Section §2.1), and Figure 4 shows the distribution across all compounds in comparison to the distributions across compounds with the same example heads as above, *Spiel*, *Werk* and *Zug*. In Figure 3 we can see that we have similar overall concreteness distributions for the compounds, the modifiers and the heads. When zooming into compounds with specific heads in Figure 4, we observe a more diverse picture: while the compounds, modifiers and heads of *Spiel* and *Zug* compounds are again skewed towards concreteness, the compounds and constituents of *Werk* compounds exhibit more diversity in their concreteness ratings.

Figure 5 and Table 6 look into the compositionality ratings in our novel dataset, making use of two perspectives. Figure 5 shows boxplots of compound–modifier and compound–head compositionality ratings. For both constituent types we can see skewed distributions towards strongly compositional compounds, similarly to the distributions in GHOST-NN, cf. Figure 1. Table 6 compares the novel ratings against the original ratings in the datasets CONCRETE-NN and GHOST-NN, relying on Spearman’s rank-order correlation coefficient ρ . The correlations are between 0.663 and 0.792 and therefore all point towards strong agreement between the novel mean ratings and the original mean ratings. On the one hand, this allows us to judge our novel collection as reliable, even though a smaller number of annotators was involved; on the other hand, the strong correlations tell us that the additional rating tasks we asked the annotators to perform did not have a strong influence on their compositionality judgements.

8 Collecting and investigating features of compositionality ratings

Table 5: Examples of mean hypernymy ratings in FEATURE-NN for a subset of the target compounds with heads *Spiel* ‘game’, *Werk* ‘work’; ‘factory’ and *Zug* ‘train’; ‘draught’.

<i>Angriffsspiel</i> ‘offensive play’	3.2	<i>Mundwerk</i> ‘gab’	0.2
<i>Ballspiel</i> ‘ball game’	4.8	<i>Netzwerk</i> ‘network’	0.4
<i>Computerspiel</i> ‘computer game’	4.8	<i>Stahlwerk</i> ‘steel plant’	5.0
<i>Farbenspiel</i> ‘play of colours’	3.0	<i>Stockwerk</i> ‘floor’	0.0
<i>Gedankenspiel</i> ‘intellectual game’	1.6	<i>Tagewerk</i> ‘day’s work’	4.0
<i>Glockenspiel</i> ‘chimes’	2.2	<i>Teufelswerk</i> ‘devil’s work’	2.8
<i>Glücksspiel</i> ‘gambling’	4.4	<i>Triebwerk</i> ‘power unit’	3.2
<i>Kartenspiel</i> ‘card game’	5.0	<i>Uhrwerk</i> ‘clockwork’	2.4
<i>Kinderspiel</i> ‘children’s game’; ‘easy’	3.6	<i>Wunderwerk</i> ‘miracle’	3.8
<i>Kirchspiel</i> ‘parish’	0.8	<i>Zementwerk</i> ‘cement plant’	4.8
<i>Liebespiel</i> ‘amorous play’	2.6	<i>Atemzug</i> ‘breath’	0.8
<i>Machtspiel</i> ‘power game’	3.2	<i>Autozug</i> ‘car train’	5.0
<i>Orgelspiel</i> ‘organ playing’	4.0	<i>Beutezug</i> ‘foray’	3.2
<i>Ritterspiel</i> ‘knights game’	4.0	<i>Charakterzug</i> ‘character trait’	2.6
<i>Schattenspiel</i> ‘shadow play’	4.2	<i>Dampfzug</i> ‘steam train’	5.0
<i>Trauerspiel</i> ‘fiasco’	2.9	<i>Fackelzug</i> ‘torchlight procession’	2.4
<i>Wasserspiel</i> ‘water game’	3.4	<i>Feldzug</i> ‘campaign’	1.4
<i>Windspiel</i> ‘wind chimes’	2.6	<i>Gebirgszug</i> ‘mountain range’	1.6
<i>Wortspiel</i> ‘pun’	3.2	<i>Gesichtszug</i> ‘facial feature’	1.0
<i>Würfelspiel</i> ‘game of dice’	5.0	<i>Kriegszug</i> ‘military expedition’	2.6
<i>Bergwerk</i> ‘mine’	4.6	<i>Luftzug</i> ‘draught’	2.8
<i>Blattwerk</i> ‘foliage’	1.6	<i>Nachtzug</i> ‘night train’	5.0
<i>Erstlingswerk</i> ‘first work’	3.4	<i>Protestzug</i> ‘protest march’	3.6
<i>Feuerwerk</i> ‘fireworks’	2.4	<i>Schachzug</i> ‘chess move’; ‘gambit’	2.5
<i>Hexenwerk</i> ‘sorcery’; ‘difficult’	2.8	<i>Schriftzug</i> ‘lettering’	0.6
<i>Klavierwerk</i> ‘piano work’	3.0	<i>Seilzug</i> ‘cable pull’	3.6
<i>Kraftwerk</i> ‘power station’	4.4	<i>Siegeszug</i> ‘triumphal march’	1.4
<i>Mauerwerk</i> ‘masonry’	2.0	<i>Trauerzug</i> ‘funeral procession’	3.6
<i>Meisterwerk</i> ‘masterpiece’	3.2	<i>Triumphzug</i> ‘triumphal march’	3.4
<i>Menschenwerk</i> ‘man-made’	4.0	<i>Vogelzug</i> ‘bird migration’	1.8

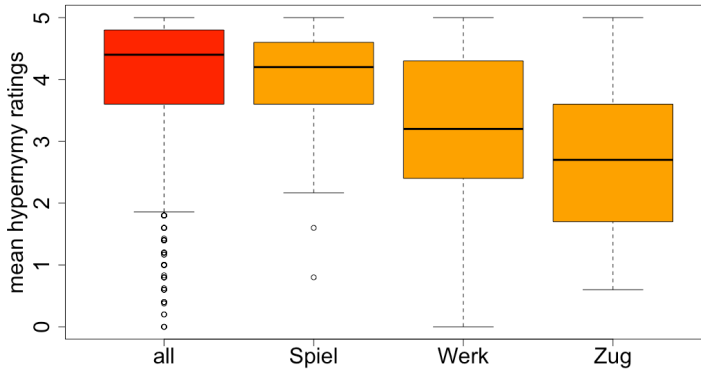


Figure 2: Strengths of hypernymy relation ratings in FEATURE-NN regarding all compound-head combinations in comparison to compounds with heads *Spiel*, *Werk* and *Zug*.

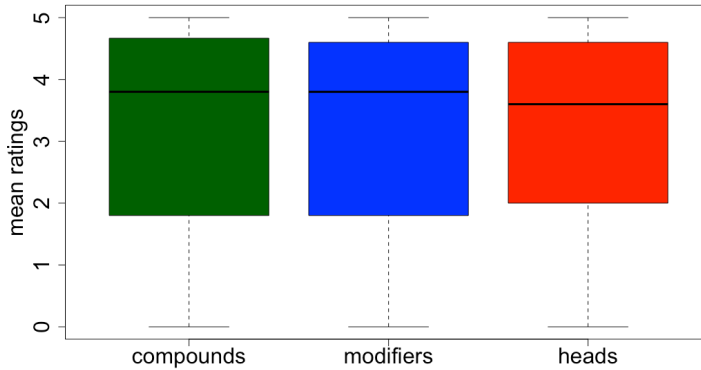


Figure 3: Concreteness ratings in FEATURE-NN.

Table 6: Correlations (ρ) between original and feature-based compositionality ratings for CONCRETE-NN and Ghost-NN compounds.

	constituent	ρ
CONCRETE-NN	modifier	0.792
	head	0.728
Ghost-NN/S	modifier	0.770
	head	0.687
Ghost-NN/XL	modifier	0.663
	head	0.687

8 Collecting and investigating features of compositionality ratings

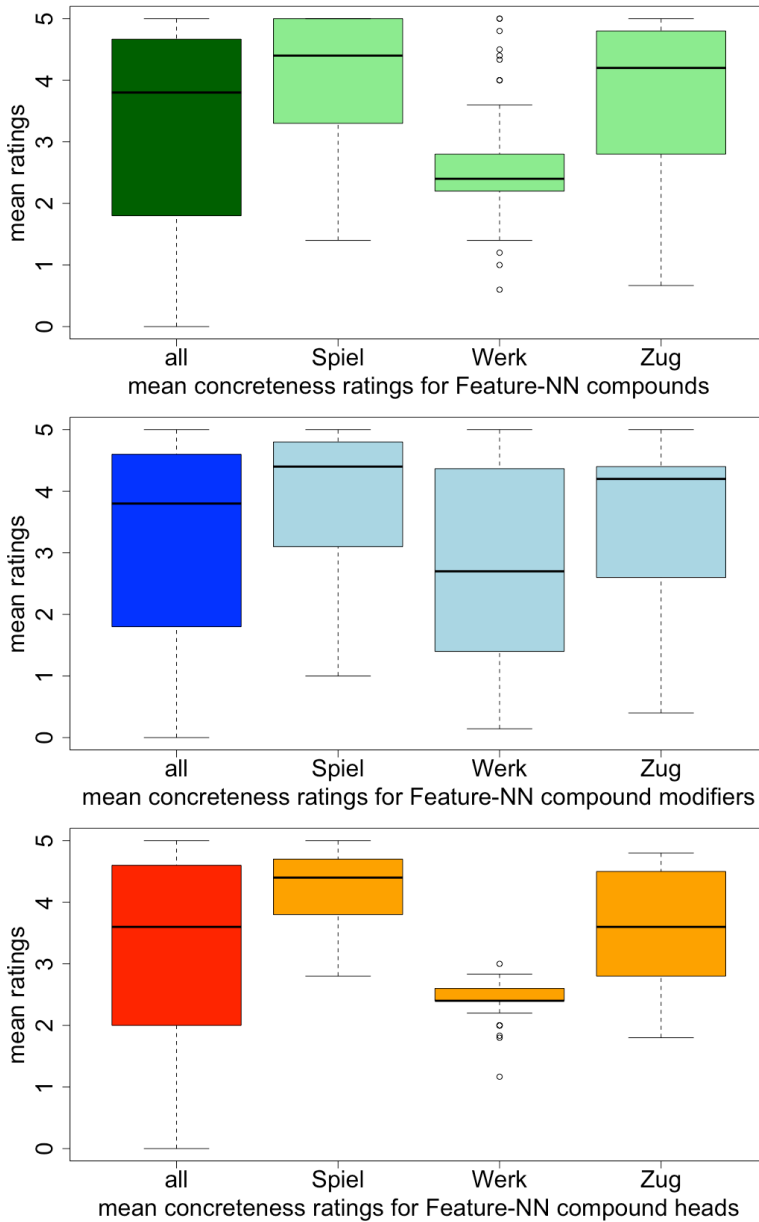


Figure 4: Concreteness ratings in FEATURE-NN, comparing ratings across all compounds (top), all compound-modifier combinations (middle), and all compound-head combinations (bottom) against those for compounds with heads *Spiel*, *Werk* and *Zug*, respectively.

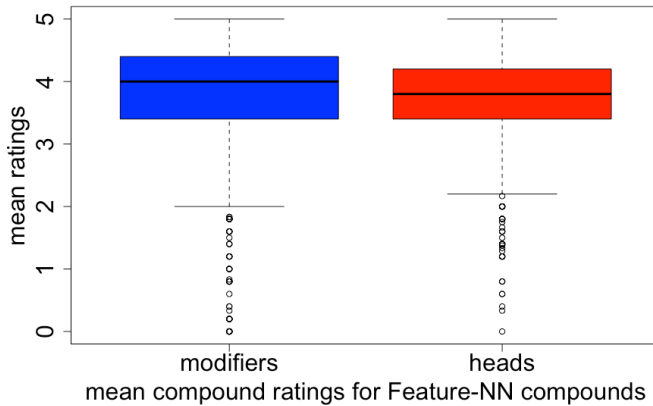


Figure 5: Compositionality ratings in FEATURE-NN.

4 Analyses

In this section, we raise and discuss two issues that we consider important for the creation of datasets with compositionality ratings, and potentially also for the creation of datasets with ratings on further semantic variables. (1) On the one hand, we are asking whether the distribution of ratings across a pre-specified scale of ratings should be even, as opposed to being skewed towards parts of the rating scale. (2) On the other hand, we are asking to what extent one should take into account properties of targets when creating a novel resource, and also when using a resource for evaluating computational models. In the following, we will look into rating distributions across datasets regarding issue (1), and into interactions between target properties and rating distributions regarding issue (2). As datasets, we will make use of the existing German and English resources CONCRETE-NN, GHOST-NN, REDDY-NN and CORDEIRO-N² introduced in Section §2.1, as well as our novel resource FEATURE-NN introduced in the previous Section §3. As properties, we will make use of frequency, productivity and ambiguity values provided by Schulte im Walde et al. (2016a) and Miletic & Schulte im Walde (2023), hypernymy and concreteness ratings for the German targets collected in FEATURE-NN, and concreteness ratings for the English compound and constituent targets collected by Muraki et al. (2022) and Brysbaert et al. (2014), respectively.

Figure 1 on page 277 presented the distributions of compositionality ratings across the targets in the existing German and English rating datasets; Figure 5 on the previous page presented the distributions for our novel dataset FEATURE-NN.

²Plots for the REDDY-NN extension CORDEIRO-N can be found in the appendix.

The two GHOST-NN variants and also our novel dataset FEATURE-NN are skewed towards strongly compositional targets, while the targets in CONCRETE-NN and even more so in REDDY-NN exhibit more even distributions. Figures 6 and 7 provide an additional view on the ratings in the latter two datasets, where the mean ratings on the x -axes are plotted in relation to the respective standard deviations (y -axes). The plots in Figures 6 and 7 confirm that there are more strongly compositional than strongly non-compositional or mid-scale targets in CONCRETE-NN, while REDDY-NN predominantly includes strongly compositional and also strongly non-compositional targets, in contrast to the mid-range which is covered rather sparsely. Overall, we induce from the distribution plots that (a) the concreteness-focused selection of targets for CONCRETE-NN, (b) the property-based balancing selection of targets for GHOST-NN, and (c) the target selection combining WordNet-based hypernymy and gloss overlap resulted in target sets with rather different distributions across compositionality ratings.

Table 7 looks into relations between compositionality ratings for compounds and compound-constituent combinations, by presenting correlations between the compositionality rating distributions for compounds and constituents within datasets. While we do not see meaningful correlations between the compound-modifier or the compound-head ratings in the GHOST-NN variants or FEATURE-NN, we find a weak negative correlation for CONCRETE-NN ($\rho = -0.372$) and weak positive correlations for REDDY-NN ($\rho = 0.265$) and CORDEIRO-N ($\rho = 0.353$). Even more so, we find strong correlations between compound and compound-modifier ratings (CONCRETE-NN: $\rho = 0.600$; REDDY-NN: $\rho = 0.804$; CORDEIRO-N: $\rho = 0.798$), and also between compound and compound-head ratings (REDDY-NN: $\rho = 0.720$ and CORDEIRO-N: $\rho = 0.759$). I.e., in CONCRETE-NN and REDDY-NN strongly compositional compounds include strongly meaning-contributing modifiers (and heads, in the datasets REDDY-NN and CORDEIRO-N), and strongly non-compositional compounds include strongly non-contributing modifiers (and heads). We will discuss these insights further after we have looked into compound properties across datasets, i.e., issue (2).

Tables 8 and 9 look into interactions between compositionality ratings and properties of compounds and constituents, again relying on Spearman's ρ correlations. More specifically, Table 8 shows correlations between compound ratings and compound frequency (freq), hypernymy (hyp), concreteness (conc), and also between compound-modifier ratings (modifier) and compound-head ratings (head) and the respective modifier/head properties, as well as productivity (prod) referring to the family size, and ambiguity (amb) referring to the number of senses. For the REDDY-NN and the CORDEIRO-N datasets, we do not have hypernymy ratings, but we assume that hypernymy is strongly involved in compound-

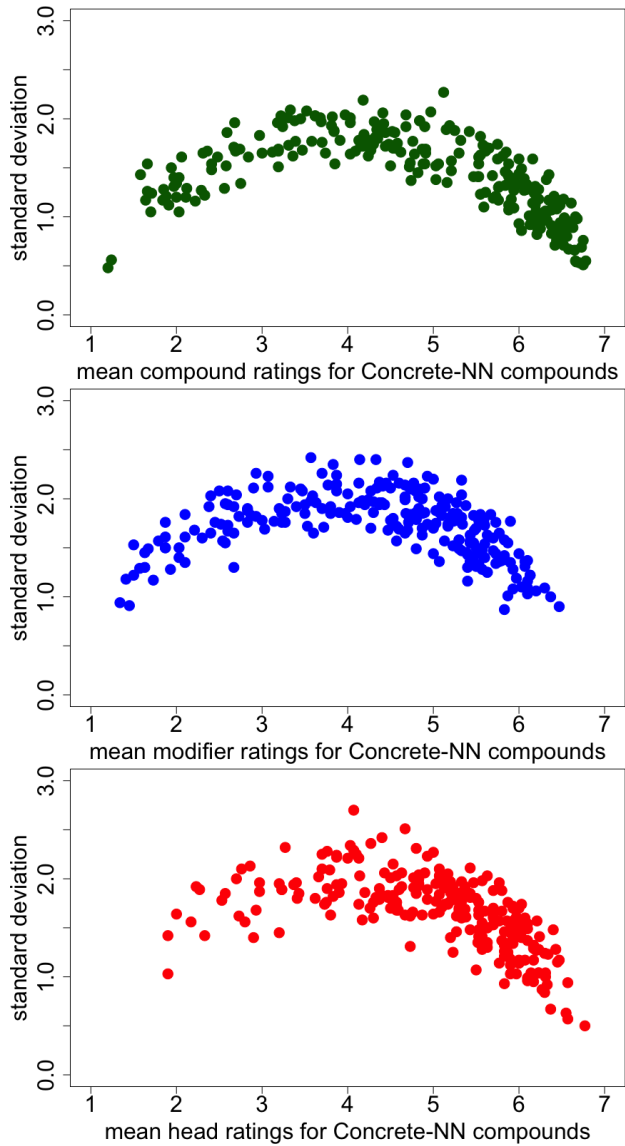


Figure 6: Mean compositionality ratings and standard deviations in CONCRETE-NN.

8 Collecting and investigating features of compositionality ratings

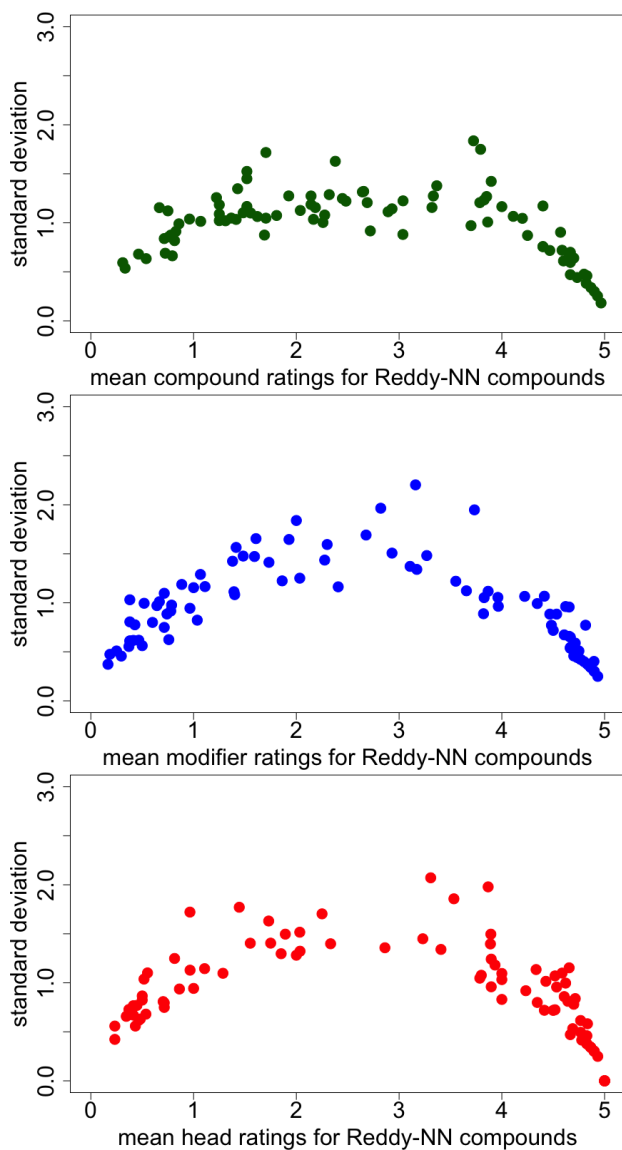


Figure 7: Mean compositionality ratings and standard deviations in REDDY-NN.

Table 7: Within-dataset correlations (ρ) between the compositionality ratings for compounds, modifiers and heads.

		ρ	
		modifier	head
German datasets			
CONCRETE-NN	compound	0.600	0.138
	modifier		-0.372
Ghost-NN/S	modifier		-0.087
Ghost-NN/XL	modifier		-0.123
Feature-NN	modifier		0.085
English datasets			
REDDY-NN	compound	0.804	0.720
	modifier		0.265
CORDEIRO-N	compound	0.798	0.759
	modifier		0.353

constituent relationships because of how targets were selected (cf. Section §2.1). We distinguish between original ratings (ORIG) and novel ratings (FEAT) in the German datasets, and we highlight cells with moderate-to-strong correlations $\rho > 0.4$.

The following observations are particularly striking: in the German dataset variants, we find a strong correlation between compound–head ratings and the degree of hypernymy ($0.624 \leq \rho \leq 0.797$), i.e., the stronger the degree of hypernymy, the more a head has been judged as contributing its meaning to the compound meaning, which we consider an indirect confirmation of the reliability of the ratings, because this is hypernymy per definitionem. In the FEATURE-NN ratings for the CONCRETE-NN compound–head combinations we further see a moderate correlation between the ratings and the heads’ degrees of concreteness ($\rho = 0.414$). For compounds, the same type of correlation is even stronger in the REDDY-NN and the CORDEIRO-N datasets ($\rho = 0.592$ and $\rho = 0.469$, respectively), and negative for the concreteness of compound–modifier ratings in REDDY-NN ($\rho = -0.492$). Most striking in the table are the moderate correlations for REDDY-NN between all compound and compound–constituent ratings and their empirical properties frequency and productivity ($0.454 \leq \rho \leq 0.579$), while there are no moderate correlations between compositionality ratings and frequency and productivity in the German datasets.

8 Collecting and investigating features of compositionality ratings

Table 8: Correlations (ρ) between compound and constituent compositionality ratings and compound and constituent properties.

			Properties				
			freq	prod	amb	hyp	conc
CONCRETE-NN	ORIG	compound	-0.075	-	-	0.424	0.113
CONCRETE-NN	ORIG	modifier	0.080	0.164	-0.157	-	0.079
CONCRETE-NN	ORIG	head	-0.147	-0.178	-0.279	0.689	0.228
CONCRETE-NN	FEAT	modifier	0.020	0.114	-0.177	0.080	0.182
CONCRETE-NN	FEAT	head	-0.070	-0.061	-0.230	0.762	0.414
Ghost-NN/S	ORIG	modifier	0.032	0.024	-0.235	-	0.002
Ghost-NN/S	ORIG	head	-0.220	-0.271	-0.305	0.797	0.344
Ghost-NN/S	FEAT	modifier	0.020	0.071	-0.192	-	0.142
Ghost-NN/S	FEAT	head	-0.164	-0.197	-0.119	0.624	0.281
Ghost-NN/XL	ORIG	modifier	-0.088	-0.023	-0.231	-	0.119
Ghost-NN/XL	ORIG	head	-0.202	-0.204	-0.356	0.692	0.171
Ghost-NN/XL	FEAT	modifier	-0.130	-0.087	-0.164	-	0.212
Ghost-NN/XL	FEAT	head	-0.246	-0.250	-0.294	0.645	0.224
REDDY-NN		compound	0.579	-	-	-	0.592
REDDY-NN		modifier	0.547	0.471	0.172	-	-0.492
REDDY-NN		head	0.454	0.484	0.224	-	-0.207
CORDEIRO-N		compound	0.385	-	-	-	0.469
CORDEIRO-N		modifier	0.340	0.269	-0.100	-	-0.381
CORDEIRO-N		head	0.307	0.331	0.110	-	-0.283

Table 9: Correlations (ρ) between compound compositionality ratings and compound and constituent properties.

	frequency			productivity		ambiguity	
	comp	mod	head	mod	head	mod	head
CONCRETE-NN	-0.075	0.049	0.099	0.101	0.199	-0.182	-0.060
REDDY-NN	0.579	0.535	0.393	0.517	0.464	0.219	0.133
CORDEIRO-N	0.385	0.188	0.257	0.132	0.314	-0.140	0.072

In Table 9, we focus on compound ratings, this time looking into correlations between compound ratings and compound and constituent properties. We can see that the compound phrase/whole ratings in the REDDY-NN dataset are also moderately correlated with modifier and head frequencies and productivities.

We now turn towards a discussion of the analyses with regard to the two issues we raised: (1) to what extent one should aim for an even distribution of ratings across the pre-specified scale of ratings, and (2) to what extent one should take into account properties of targets when creating a novel resource and when using a resource for evaluation. We saw in our analyses that the datasets we explored are skewed towards certain ranges of compositionality in different ways, some contain more compositional than non-compositional compounds, and some contain many more ratings at either extreme of the compositionality scale than in the mid-range. Furthermore, in some datasets (but not in others) we find strong correlations between compound and compound-constituent ratings as well as moderate correlations between compositionality ratings and corpus-based frequencies and productivity scores. Which of these inter-dependencies are desired, and which are artefacts created by the specific strategies of how to select compound targets for the dataset? Optimally, one should aim for ratings on a scale that are evenly distributed across targets, both overall and also with regard to salient target properties, in order to ensure full coverage of the phenomenon. This goal is very difficult to achieve, however, because we can only check on rating distributions once we have collected the ratings. We therefore suggest to pay attention to a subset of target properties that are considered most salient and influential regarding the desired rating types. This was done for GHOST-NN by Schulte im Walde et al. (2016a), for example, whose resulting ratings are however highly skewed towards compositionality, so in retrospect our specific choice of salient properties may be considered suboptimal.

We see two alternative routes to follow, individually or in combination: (a) Balance your targets across frequency ranges as the minimally required target property, because we know that target frequency has generally a strong influence on language processing and comprehension (Ellis 2002). (b) If time and money allow, go for a large set of targets in the selection phase, such that the collected ratings may be analysed and the targets then be post-balanced across the most salient target properties in a post-processing filtering step. Realistically, many datasets that are available or will be available in the future still incorporate artefacts with regard to one or the other target property, so we need a workaround when evaluating our computational models on the basis of such datasets. Our baseline for this workaround is to assess models not only on the full dataset, but also with regard to subsets of targets with coherent task-relevant properties, similarly to our

studies described in Section §2.2 (Schulte im Walde et al. 2016a, Köper & Schulte im Walde 2017, Alipoor & Schulte im Walde 2020, Miletic & Schulte im Walde 2023). In this way we obtain a fine-tuned set of model results, rather than “just” an overall result score.

5 Conclusion

The current study started off with the observation that evaluations of computational models predicting degrees of compositionality for noun compounds typically evaluate their models across all targets, disregarding the fact that prediction models might vary according to properties of the targets within the gold standard resources. We suggested a novel route to assess the interactions between compound and constituent properties with regard to degrees of compositionality: (1) We created a novel collection FEATURE-NN with compositionality ratings for 1,099 German compounds, where we asked the human judges to provide compound and constituent properties (such as paraphrases, meaning contributions, hypernymy relations, and concreteness) before judging the compositionality; and (2) We performed a series of analyses on rating distributions and interactions with compound and constituent properties for our novel collection as well as previous gold standard resources for German (CONCRETE-NN and GHOST-NN) and English (REDDY-NN and CORDEIRO-N). Our novel collection of ratings provides useful materials to investigate the meanings of the 1,099 compound targets and their constituents and is available from <http://www.ims.uni-stuttgart.de/data/feature-comp-nn> under a CC BY-NC-SA license. The obtained compositionality ratings are strongly correlated with previous ratings on the same targets, from which we induce (a) that we judge our novel ratings as reliable, and at the same time (b) that the additional ratings on compound and constituent properties that we asked the human judges to provide did not have a strong influence on their judgements.

Making use of our novel annotations as well as information on frequencies, productivities, ambiguities and degrees of concreteness regarding the target compounds and their constituents, we gained insight into distributions over compositionality ratings as well as interactions between these distributions and a range of target properties, most importantly: (a) The previous and also our novel collection of compositionality ratings all show skewed distributions, however in various ways: GHOST-NN and FEATURE-NN are skewed towards strongly compositional targets, while REDDY-NN includes strongly compositional and also strongly non-compositional targets while the mid-range is covered more sparsely.

(b) Regarding relations between compound and constituent ratings, CONCRETE-NN and REDDY-NN show moderate-to-strong correlations between compound and compound-modifier ratings (CONCRETE-NN: $\rho = 0.600$; REDDY-NN: $\rho = 0.804$) and between compound and compound-head ratings (REDDY-NN: $\rho = 0.720$). (c) Looking into the interactions between compound and constituent properties and their compositionality ratings, we found moderate-to-strong correlations with concreteness (CONCRETE-NN: $\rho = 0.414$; and REDDY-NN: $\rho = 0.592$ for compounds and $\rho = 0.492$ for heads), and we also found moderate correlations with frequency and productivity (REDDY-NN: $0.393 \geq \rho \geq 0.579$).

Following the analyses we discussed to what extent one should aim for an even distribution of ratings across the pre-specified scale, and to what extent one should take into account properties of targets when creating a novel resource and when using a resource for evaluation. We suggest as a minimum requirement to balance targets across frequency ranges, and optimally to balance targets across their most salient properties in a post-collection filtering step. Above all, we recommend assessing computational models not only on the full dataset but also with regard to subsets of targets with coherent task-relevant properties. We believe that especially the latter recommendation does not only apply to compositionality ratings (resources and models) but more generally to creating and using evaluation datasets across tasks.

Abbreviations

MWE	multiword expression
NLP	natural language processing
NLU	natural language understanding
BE	semantic compound relation: be
HAVE	semantic compound relation: have
IN	semantic compound relation: in
ABOUT	semantic compound relation: about
ACTOR	semantic compound relation: actor
INST	semantic compound relation: instrument
LEX	no semantic compound relation; lexicalised compound

Acknowledgements

We thank the five annotators for their contributions to the creation of the dataset FEATURE-NN, and we thank Chris Jenkins and Filip Miletic as well as the two

anonymous reviewers and the editors of this volume for their feedback on previous versions of this chapter. Our research received funding from the German Research Foundation (DFG) through projects *Sense Discrimination and Regular Meaning Shifts of German Particle Verbs* in the Collaborative Research Centre SFB 732, SCHU 2580/5 *Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*, and SCHU 2580/2 *Distributional Approaches to Semantic Relatedness*.

Appendix A Annotation guidelines for FEATURE-NN ratings

A.1 Original German version: Guidelines für die Annotation von Eigenschaften komplexer Nomen und ihrer Konstituenten

In der Datei `anno-comp-ratings-feat.ods` findest Du eine Liste von komplexen Nomen und ihren zwei nominalen Konstituenten in den Spalten A, B und C (und für eine bessere Übersichtlichkeit wiederholt in den Spalten M, N und O). In den dazwischen liegenden Spalten bitten wir Dich um Deine spontanen Intuitionen bezüglich folgender Eigenschaften:

Spalte D: **Bedeutung des komplexen Nomens**

Aufgabe: Erkläre die Bedeutung des komplexen Nomens in einer Phrase/einem Satz. Du darfst (musst aber nicht) die Konstituenten des Nomens in Deiner Erklärung verwenden.

Beispiel: Die Bedeutung des komplexen Nomens *Eselsohr* ist *verknickte Ecke einer Buchseite*.

Spalten E und F: **Eigenschaften der Konstituenten**

Welche Eigenschaften der ersten bzw. zweiten Konstituente finden sich in dem komplexen Nomen wieder? Falls Dir mehrere Eigenschaften einfallen, trenne diese bitte durch Komma. Falls Dir keine Eigenschaft einfällt, trage bitte "0" ein.

Beispiel: Bei dem komplexen Nomen *Kunstfehler* trägt z.B. die erste Konstituente die Eigenschaften *sehr gut*, *Qualität* bei, die zweite Konstituente z.B. die Eigenschaft *Misserfolg*.

Versuche, jede Eigenschaft auf ein oder wenige Worte zu beschränken. Die Wortarten sind beliebig.

Spalte G: **Über-/Unterordnung**

Ist das komplexe Nomen “eine Art” der zweiten Konstituente? Nutze eine Skala von 0 (nein, gar nicht) bis 5 (ja, absolut).

Beispiel: “Ein Ahornbaum ist eine Art von Baum”, aber
“Ein Eselsohr ist **keine** Art von Ohr”.

Spalten H–J: **Abstraktheit/Konkretheit**

Wie abstrakt bzw. konkret sind das komplexe Nomen sowie die erste bzw. zweite Konstituente? Nutze wiederum eine Skala von 0 (ganz abstrakt) bis 5 (ganz konkret).

Hinweis: Konkrete Wörter können durch die menschlichen Sinne (hören, riechen, schmecken, sehen, tasten) erfasst werden (z.B. *Tisch, Lärm*), abstrakte Wörter nicht (z.B. *Idee, Traum*).

Spalten K–L: **Kompositionalität**

Wie sehr lässt sich die Gesamtbedeutung des komplexen Nomens aus der Bedeutung der ersten bzw. zweiten Konstituente ableiten? Nutze wiederum eine Skala von 0 (gar nicht) bis 5 (sehr stark).

A.2 Tentative English translation: Guidelines for annotating properties of complex nouns and their constituents

The file `anno-comp-ratings-feat.ods` provides a list of complex nouns and their two nominal constituents in columns A, B and C (and repeated in columns M, N and O). In the intermediate columns we ask for your spontaneous intuitions regarding the following properties:

Column D: **Meaning of the complex noun**

Task: Explain the meaning of the complex noun within one phrase/sentence. You may (but you do not have to) use the constituents of the noun in your explanation.

Example: The meaning of the complex noun *Eselsohr* (lit. ‘donkey ear’) ‘*earmark*’ is a *folded corner of a page in a book*.

Columns E and F: **Properties of the constituents**

Which properties of the first/second constituent do you recognise in the complex noun? If you are aware of several properties, please separate them with commas. If you are not aware of any property, please enter “0”.

8 Collecting and investigating features of compositionality ratings

Example: Regarding the complex noun *Kunstfehler* (lit. ‘art mistake’) ‘mal-practice’ the first constituent contributes the properties *excellent* and *quality*, and the second constituent contributes the property *failure*.

Try to use only one or a few words for each property. You may use words of any word class.

Column G: Super-/subordination

Is the complex noun “a kind of” the second constituent? Please use a scale between 0 (no, not at all) and 5 (yes, absolutely).

Example: “An *Ahornblatt* ‘maple tree’ is a kind of tree”, but
“An *Eselsohr* (lit. ‘donkey ear’) ‘earmark’ is **not** a kind of ear”.

Columns H–J: Abstractness/concreteness

How abstract/concrete are the complex noun and the first and second constituent? Again, please use a scale between 0 (totally abstract) and 5 (totally concrete).

Hint: Concrete words can be perceived by human senses: hearing, smelling, tasting, seeing, touching (e.g., *table*, *noise*), abstract words cannot (e.g., *idea*, *dream*).

Columns K–L: Compositionality

To what degree can you induce the meaning of the complex nouns from the meanings of the first/second constituents? Again, please use a scale between 0 (not at all) and 5 (totally).

Appendix B Cordeiro dataset ratings

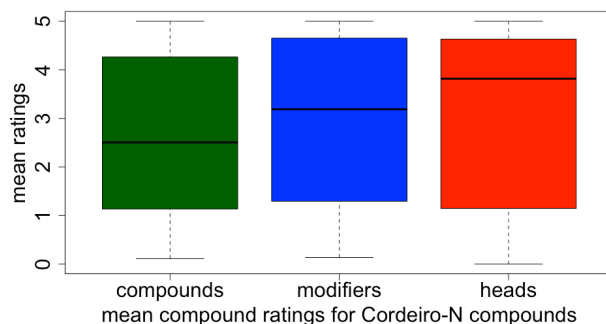


Figure 8: Compositionality rating distributions in CORDEIRO-N.

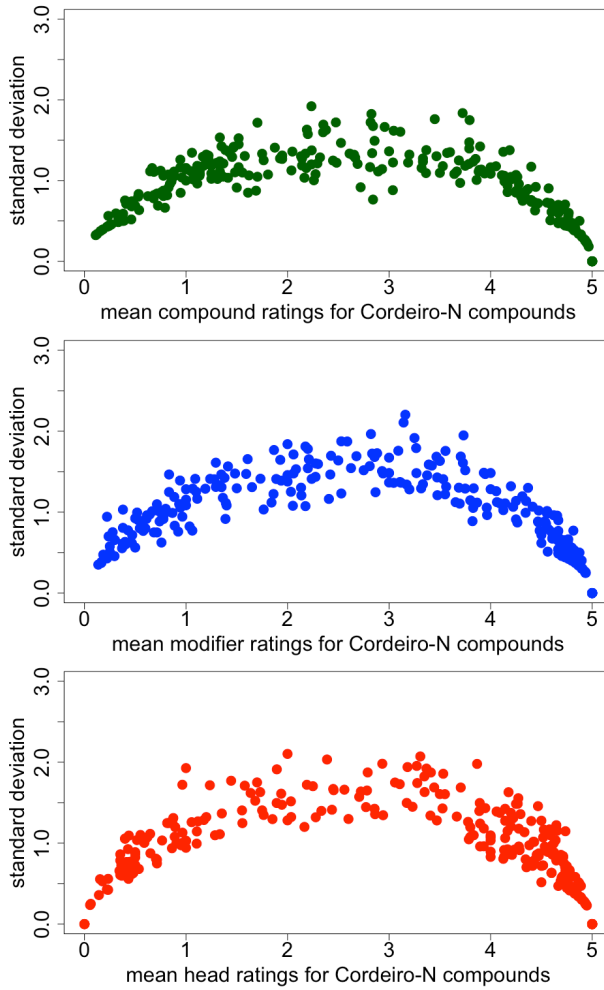


Figure 9: Mean compositionality ratings and standard deviations for compounds in CORDEIRO-N.

Appendix C Concreteness of Targets in REDDY-NN and CORDEIRO-N

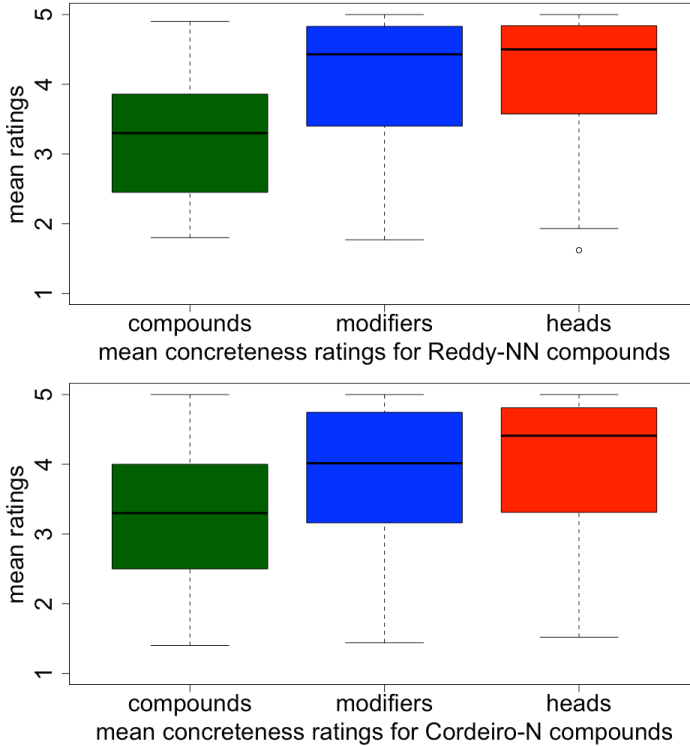


Figure 10: Concreteness ratings in REDDY-NN and CORDEIRO-N.

References

Alipoor, Pegah & Sabine Schulte im Walde. 2020. Variants of vector space reductions for predicting the compositionality of English noun compounds. 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 (eds.), *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'2020)*, 4379–4387. Marseille, France: ACL. <https://aclanthology.org/2020.lrec-1.539>.

- Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 267–292. Boca Raton, FL: CRC Press.
- Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.
- Bauer, Laurie. 2017. *Compounds and compounding*. Cambridge University Press.
- Bell, Melanie J. & Martin Schäfer. 2013. Semantic transparency: Challenges for distributional semantics. In Aurelie Herbelot, Roberto Zamparelli & Gemma Boleda (eds.), *Proceedings of the IWCS 2013 workshop on formal distributional semantics*, 1–10. Potsdam, Germany.
- Benczes, Réka. 2014. What can we learn about the mental lexicon from non-prototypical cases of compounding? *Argumentum* 10. 205–220.
- Bettinger, Julia, Anna Hättö, Michael Dorna & Sabine Schulte im Walde. 2020. A domain-specific dataset of difficulty ratings for German noun compounds in the domains DIY, cooking and automotive. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'2020)*, 4359–4367. Marseille, France: European Language Resources Association (ELRA).
- Bott, Stefan & Sabine Schulte im Walde. 2017. Factoring ambiguity out of the prediction of compositionality for German multi-word expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, 66–72. Valencia, Spain. DOI: 10.18653/v1/W17-1708.
- Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 64. 904–911. DOI: 10.3758/s13428-013-0403-5.
- Butterworth, Brian. 1983. Lexical representation. In *Language production*, vol. 2: Development, writing and other language processes, 257–294. London: Academic Press.
- Cap, Fabienne, Manju Nirmal, Marion Weller & Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th workshop on multiword expressions*, 19–28. Denver, CO.
- Carpuat, Marine & Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Proceedings of the 11th annual conference of the North American chapter of the Association for Computational Linguistics*. Los Angeles, CA.

- Cholakov, Kostadin & Valia Kordoni. 2014. Better statistical machine translation through linguistic treatment of phrasal verbs. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, 196–201. Doha, Qatar.
- Clouet, Elizaveta Loginova & Béatrice Daille. 2014. Splitting of compound terms in non-prototypical compounding languages. In Ben Verhoeven, Walter Daelemans, Menno van Zaanen & Gerhard van Huyssteen (eds.), *Proceedings of the 1st workshop on Computational Approaches to Compound Analysis (ComACoMA 2014)*, 11–19. Dublin, Ireland: ACL. DOI: 10.3115/v1/W14-5702.
- Cordeiro, Silvio, Aline Villavicencio, Marco Idiart & Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics* 45(1). 1–57.
- Costello, Fintan J. & Mark T. Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science* 24(2). 299–349.
- Dankers, Verna, Elia Bruni & Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In Smaranda Muresan, Preslav Nakov & Aline Villavicencio (eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics*, 4154–4175. Dublin, Ireland. DOI: 10.18653/v1/2022.acl-long.286.
- de Jong, Nicole H., Laurie B. Feldman, Robert Schreuder, Michael Pastizzo & R. Harald Baayen. 2002. The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language* 81. 555–567.
- Dima, Corina, Daniel de Kok, Neele Witte & Erhard Hinrichs. 2019. No word is an island: A transformation weighting model for semantic composition. *Transactions of Computational Linguistics* 7. 437–451.
- Eichel, Annerose, Helena Schlipf & Sabine Schulte im Walde. 2023. *Made of Steel?* Learning plausible materials for components in the vehicle repair domain. In *Proceedings of the 17th conference of the European chapter of the Association for Computational Linguistics*, 1420–1435. Dubrovnik, Croatia.
- Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2). 143–188.
- Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database* (Language, Speech, and Communication). Cambridge, MA: MIT Press.
- Firth, John R. 1957. *Papers in linguistics 1934–1951*. London, UK: Longmans.
- Gagné, Christina L. 2002. Lexical and relational influences on the processing of novel compounds. *Brain and Language* 81. 723–735.

- Gagné, Christina L., Thomas L. Spalding & Daniel Schmidtke. 2019. LADEC: The large database of English compounds. *Behavior Research Methods* 51. 2152–2179.
- Gagné, Christina L., Thomas L. Spalding, Patricia Spicer, Dixie Wong & Beatriz Rubio. 2020. Is *buttercup* a kind of cup? Hyponymy and semantic transparency in compound words. *Journal of Memory and Language* 113. DOI: 10.1016/j.jml.2020.104110.
- Gamallo, Pablo, Susana Sotelo, Jose Ramon Pichel & Mikel Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics* 45(3). 395–421.
- Girju, Roxana, Dan Moldovan, Marta Tatu & Daniel Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language* 19(4). 479–496.
- Hamp, Birgit & Helmut Feldweg. 1997. GermaNet: A lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*. <https://aclanthology.org/W97-0802>.
- Harris, Zellig. 1954. Distributional structure. *Word* 10(23). 146–162.
- Hätty, Anna, Julia Bettinger, Michael Dorna, Jonas Kuhn & Sabine Schulte im Walde. 2021. Compound or term features? Analyzing salience in predicting the difficulty of German noun compounds across domains. In *Proceedings of the 10th joint conference on lexical and computational semantics, 252–262*. Bangkok, Thailand.
- Hätty, Anna, Ulrich Heid, Anna Moskvina, Julia Bettinger, Michael Dorna & Sabine Schulte im Walde. 2019. AkkuBohrHammer vs. AkkuBohrhammer: Experiments towards the evaluation of compound splitting tools for general language and specific domains. In *Proceedings of the 15th conference on natural language processing (KONVENS 2019)*, 59–67. Erlangen, Germany: German Society for Computational Linguistics & Language Technology.
- Hätty, Anna & Sabine Schulte im Walde. 2018. Fine-grained termhood prediction for German compound terms using neural networks. In *Proceedings of the COLING joint workshop on linguistic annotation, multiword expressions and constructions*, 62–73. Santa Fe, NM.
- Hermann, Karl Moritz. 2014. *Distributed representations for compositional semantics*. University of Oxford. (Doctoral dissertation).
- Köper, Maximilian & Sabine Schulte im Walde. 2017. Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th workshop on multiword expressions*, 200–206. Valencia, Spain.

- Kunze, Claudia. 2000. Extension and use of GermaNet, a lexical-semantic database. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.), *Proceedings of the 2nd international Conference on Language Resources and Evaluation (LREC'00)*, 999–1002. Athens, Greece: European Language Resources Association (ELRA). <https://aclanthology.org/L00-1274/>.
- Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. London: Academic Press.
- Libben, Gary, Martha Gibson, Yeo Bom Yoon & Dominiek Sandra. 1997. Semantic transparency and compound fracture. *CLASNET Working Papers* (9). 1–13.
- Libben, Gary, Martha Gibson, Yeo Bom Yoon & Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84. 50–64.
- Marsh, Charles. 2015. *Cigarette helmets and horse wars: Towards a better understanding of noun compound interpretability*. Department of Computer Science, Princeton University. (Bachelor thesis).
- Miletic, Filip & Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th conference of the European chapter of the Association for Computational Linguistics*, 1499–1512. Dubrovnik, Croatia.
- Mitchell, Jeff & Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34. 1388–1429.
- Muraki, Emiko J., Summer Abdalla, Marc Brysbaert & Penny M. Pexman. 2022. Concreteness ratings for 62 thousand English multiword expressions. DOI: 10.31234/osf.io/m397u.
- Murphy, Gregory L. 1990. Noun phrase interpretation and conceptual combination. *Journal of Memory and Language* 29. 259–288.
- Nastase, Viviana A. 2003. *Semantic relations across syntactic levels*. School of Information Technology & Engineering, University of Ottawa. (Doctoral dissertation).
- Ó Séaghdha, Diarmuid. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics (CL2007)*. Birmingham, UK: University of Birmingham, UK. <https://ucrel.lancs.ac.uk/publications/cl2007/>.
- Plag, Ingo. 2003. *Word-formation in English*. Cambridge University Press.
- Reddy, Siva, Ioannis P. Klapaftis, Diana McCarthy & Suresh Manandhar. 2011a. Dynamic and static prototype vectors for semantic composition. In *Proceedings of the 5th international joint conference on Natural Language Processing*, 705–713. Chiang Mai, Thailand.

- Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011b. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th international joint conference on natural language processing*, 210–218. Chiang Mai, Thailand.
- Roller, Stephen & Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the conference on empirical methods in natural language processing*, 1146–1157. Seattle, WA, USA.
- Roller, Stephen & Sabine Schulte im Walde. 2014. Feature norms of German noun compounds. In *Proceedings of the 10th workshop on multiword expressions*, 104–108. Gothenburg, Sweden.
- Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander F. Gelbukh (ed.), *Proceedings of the third international conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, 1–15. Springer.
- Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Proceedings of the 2nd joint conference on lexical and computational semantics*, 266–275. Atlanta, GA.
- Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014a. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the conference on empirical methods in Natural Language Processing*, 1792–1797. Doha, Qatar.
- Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014b. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In Shuly Wintner, Sharon Goldwater & Stefan Riezler (eds.), *Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics*, 472–481. Gothenburg, Sweden: ACL. DOI: 10.3115/v1/E14-1050.
- Salehi, Bahar, Paul Cook & Timothy Baldwin. 2015a. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics/human language technologies*, 977–983. Denver, Colorado, USA.
- Salehi, Bahar, Nitika Mathur, Paul Cook & Timothy Baldwin. 2015b. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th workshop on multiword expressions*, 54–59. Denver, CO.
- Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejcek, Fabienne Cap, Slavomir Ceplo, Silvio Ricardo Cordeiro, Gulsen Eryigit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartin, Lonneke van der

- Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.14715.
- Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lungen & Andreas Witt (eds.), *Proceedings of the 3rd workshop on Challenges in the Management of Large Corpora (CMLC-3)*, 28–34. Mannheim, Germany.
- Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/pdf/834_Paper.pdf.
- Schulte im Walde, Sabine & Susanne Borgwaldt. 2015. Association norms for German noun compounds and their constituents. *Behavior Research Methods* 47(4). 1199–1221.
- Schulte im Walde, Sabine, Anna Häty & Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the 5th joint conference on lexical and computational semantics*, 148–158. Berlin, Germany.
- Schulte im Walde, Sabine, Anna Häty, Stefan Bott & Nana Khvtisavrishvili. 2016b. G_nost-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the 10th international conference on Language Resources and Evaluation*, 2285–2292. Portoroz, Slovenia.
- Schulte im Walde, Sabine, Stefan Müller & Stephen Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun Compounds. In Mona Diab, Tim Baldwin & Marco Baroni (eds.), *Proceedings of the 2nd joint conference on Lexical and Computational Semantics*, 255–265. Atlanta, GA. <https://aclanthology.org/S13-1038>.
- Schulte im Walde, Sabine & Eva Smolka (eds.). 2020. *The role of constituents in multi-word expressions: An interdisciplinary, cross-lingual perspective* (Phraseology and Multiword Expressions 4). Berlin: Language Science Press. DOI: 10.5281/zenodo.3598577.

- Siegel, Sidney & N. John Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. Boston, MA: McGraw-Hill.
- Spalding, Thomas L., Christina L. Gagné, A. C. Mullaly & Hongbo Ji. 2010. Relation-based interpretation of noun-noun phrases: A new theoretical approach. In Susan Olsen (ed.), *New impulses in word-formation* (Linguistische Berichte Sonderhefte 17), 283–316.
- Taft, Marcus & Kenneth I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14. 638–648.
- von der Heide, Claudia & Susanne Borgwaldt. 2009. Assoziationen zu Unter-, Basis- und Oberbegriffen: Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, 51–74.
- Weller, Marion, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde & Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the 1st workshop on computational approaches to compound analysis*, 81–90. Dublin, Ireland.
- Wisniewski, Edward J. 1996. Construal and similarity in conceptual combination. *Journal of Memory and Language* 35. 434–453.

Chapter 9

Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results

Therese Lindström Tiedemann^a, David Alfter^b, Yousuf Ali Mohammed^b, Daniela Piipponen^a, Beatrice Silén^a & Elena Volodina^b


^aUniversity of Helsinki, Finland ^bUniversity of Gothenburg, Sweden

This chapter introduces part of the Swedish L2 profiles, a new resource for Swedish as a second language. Multiword expressions (MWEs) in this resource are based on knowledge-based automatic annotation of MWEs, which we show works quite well for Swedish. In contrast, manual annotation of the compositionality of each MWE proved difficult, probably due to different interpretations of “compositionality” by the two annotators. We show that experts and non-experts can rank MWEs very similarly according to relative receptive difficulty, with particularly high agreement for the easiest items. A qualitative comparison of the proficiency levels associated with the MWEs based on coursebook occurrences and the results from crowdsourcing and direct ranking indicate that MWEs which appear in few books of the same level are more likely to be difficult to associate with an appropriate level based on coursebook corpus data. Furthermore, results show that compositionality and/or transparency might influence the relative ranking. Finally, there is a clear increase in MWE lemmas at higher proficiency levels at the group level, and at the highest level receptive and productive data include the same percentage of MWEs.

1 Introduction

Previous research has clearly shown that multiword expressions (MWEs) are an important part of idiomatic language use (e.g. Paquot 2019), but also that they



Therese Lindström Tiedemann, David Alfter, Yousuf Ali Mohammed, Daniela Piipponen, Beatrice Silén & Elena Volodina. 2024. Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 309–348. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998647 

are a challenge even to advanced second language (L2) learners (Pawley & Syder 1983, Wray 2002) and show a clear correlation to one's level of proficiency (Forsberg 2010). MWEs can be seen as:

...sequences of words that are in some regard *not entirely predictable*, whether on account of a meaning that is wildly or subtly different from the words they contain, a function that is only achieved with the whole expression, or features of structure such as morphology or word order that are non-canonical. (Wray 2013: 317, italics added)

This clearly entails that MWEs are an additional complication in second language acquisition (SLA). It also means that similarly to single word lexemes, MWEs have to be learnt as lexemes.

Based on previous research showing the challenges of MWEs in SLA and in relation to current advances in automatic evaluation, it is important to consider whether MWEs can be seen as particularly criterial of certain levels, but also how well MWEs can be automatically annotated in learner texts even though these texts tend to contain issues which do not follow the norm in the target language. Furthermore, ways of linking MWEs to proficiency levels need to be explored.

We are primarily interested in MWEs in relation to the acquisition of Swedish as a second language (L2 Swedish), however most of our results should be of interest also in relation to other languages and for SLA in general. Second language acquisition of Swedish MWEs has been studied through experiments, questionnaires and occasionally also in learner texts (e.g. Prentice & Sköldberg 2013; Enström 1990; Abrahamsson & Hyltenstam 2009).

In this chapter we present studies of Swedish MWEs based on authentic L2 data, both receptive and productive, linked to proficiency levels according to the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). We argue for the usefulness of automatic annotation of MWEs combined with additional manual annotation, and discuss the possibilities of linking MWEs to CEFR levels based on authentic data, crowdsourcing and expert annotation. Since many languages tend to have less resources than English and because we know that there will always be new expressions which will need to be linked to proficiency levels, we want to find a cheap and reliable way of linking MWEs to levels. We therefore explore crowdsourcing with relative judgement. Since it is likely to be cheaper to use non-experts and because it is interesting as a research question, we want to see if experts (L2 Swedish teachers, assessors, researchers) and non-experts (L2 Swedish learners) agree on their ratings. In addition, we also explore the possibility of explicit level ranking by experts. In this

chapter, we summarise previously published results (see Alfter et al. 2021, Lindström Tiedemann et al. 2022) and present further qualitative analyses of some of our results from these experiments.

Our study is based on MWEs found through automatic annotation of coursebooks for L2 Swedish aimed at adult learners (the Coctail corpus, Volodina et al. 2014) and L2 Swedish learner essays (the SweLL-pilot corpus, Volodina et al. 2016a). We summarise the results of our annotation check as previously published (Volodina et al. 2022b), showing that MWEs in these different materials can be well annotated automatically.

The identified MWEs were manually categorised according to our Swedish taxonomy for MWEs, see Section 3. By using our taxonomy to compare receptive and productive usage of MWEs we present how our data, including our manual annotation, can be accessed through an open lexical resource online (Swedish L2 Profiles)¹ that can be used for further research, as well as for teaching.

We aim to answer the following research questions:

1. How well can MWEs be automatically annotated in L2 coursebooks and L2 learner texts?
2. How well does the occurrence of MWEs in authentic materials coincide with (a) ranking results from an expert or a non-expert crowd; (b) direct annotation by experts?
3. How do different MWE types appear over CEFR levels in receptive and productive data for L2 Swedish? Are certain MWE types more challenging to L2 Swedish learners based on a comparison of their occurrence in receptive and productive data?

First we present some previous research on MWEs in relation to SLA and L2 Swedish in particular (Section 2). We then present our Swedish MWE taxonomy (Section 3) after which our materials and method are introduced, including the annotation tools that we use (Section 4). In Section 5 we present our results, and in Section 6 we summarise our conclusions and look ahead.

2 Previous research

MWEs are a broad and vaguely defined phenomenon. Research articles and books contain a multitude of different terms with similar meanings: collocations (Bhalla

¹<https://spraakbanken.gu.se/larkalabb/svlp> (login: demo)

& Klimcikova 2019), phraseological units (Paquot 2019), lexicalised phrases (Sag et al. 2002), fixed expressions (Villada Moirón 2005), formulaic language (Durrant 2018), lexical bundles (Granger 2018), words-with-spaces (Sag et al. 2002), formulaic sequences (Wray & Perkins 2000, Wray 2002). Wray (2002) lists c. 60 terms for similar concepts and notes the problem of the varying terminology and that terms tend to have strong connections to certain theories or methods. This also means that even when the same term is used we cannot be certain that it means the same. When working on MWEs in a language other than English this causes additional challenges even with a language as closely related to English as Swedish, since this plethora of terms needs to be compared to terminology which has been used in descriptions in that language.

The multitude of terms is partly a result of the many different approaches to MWEs. Bhalla & Klimcikova (2019) name three main approaches to studying and classifying collocations, and by extension MWEs, namely:

1. Psychological – lexical associations in the mental lexicon;
2. Phraseological – dealing predominantly with separating MWEs from free word combinations based on semantic principles; and
3. Distributional – focusing on the manifestations of MWEs in corpora based on frequency, distribution, and degree of co-occurrence.

This captures the complexity of the phenomenon and the variety of ways in which it can be perceived. It also underscores the practical needs to identify MWEs and to categorise them into subcategories. There is a need to explain their (typical and atypical) behaviour in relation to various fields, e.g. lexicography, language learning, clinical linguistics and Natural Language Processing (NLP). For lexicography, it is important to have an approach to listing MWEs, to grouping them into specialised lexicons, as well as an approach for the identification of new MWEs (Agirre et al. 2006). Identification of MWEs relies on NLP approaches (Baldwin & Bond 2002, Sag et al. 2002, Piao et al. 2005, Attia et al. 2010, de Caseli et al. 2010, Watrin & François 2011, Shigeto et al. 2013) and thus requires formalisation of the definition of MWEs, something we explore further in relation to our study below.

2.1 MWEs in SLA research

Several studies have shown that MWEs are a major part of our lexical competence. Jackendoff (1997) claims that c. 50% of our mental lexicon consists of

MWEs, while Erman (2007: 28) argues that they may in fact form an even larger part of our language since Mel'čuk (1998: 24) claims that MWEs (or phrasemes) "outnumber words roughly ten to one" (where, by "words", presumably, Mel'čuk means single lexical items). Interestingly, experiments have shown that L2 and L1 speakers process language differently. L2 speakers apparently rely primarily on frequency, whereas L1 speakers rely more on Mutual Information (MI) in processing MWEs (Ellis 2012: 24).

Some researchers have claimed that MWEs are frequent also in learner language, sometimes assuming that they are more common at lower proficiency levels (Wray 2002: 173 citing the work of others). Ellis (2012: 18) claims that "Zipf's (1935) law and the 'phrasal teddy bear' explain the paradox whereby formulas seed language acquisition and yet learners typically do not achieve native-like formulaicity". Formulaic language has been claimed "the biggest stumbling block to sounding nativelike" (Wray 2002: ix). Similarly, CEFR documentation claims that "idiomatic expressions and colloquialisms" are not likely to be fully acquired before C2 (Council of Europe 2009: 185, 187), although many "idiomatic expressions and colloquialisms" should be *understood* at C1 (Council of Europe 2009: 124, 143).

Research has shown both an increased use of MWEs and more native-like usage in terms of distribution, as the proficiency increases (e.g. Forsberg 2006 as cited in Ringbom 2012 for prefabs in L2 French (L1 Swedish); Forsberg & Bartning 2010 with regards to lexical formulaic sequences). Still, MWEs remain difficult even for advanced learners (Nesselhauf 2003: 237, Ringbom 2012: 496; Ekberg 2013) as do specific MWEs such as idioms and proverbs (Abrahamsson & Hyltenstam 2009, Prentice 2010) and since it is "clearly impossible to teach all (or even most) of the collocations in a language, criteria have to be set up to determine which collocations should be included in a given syllabus" (Nesselhauf 2003: 238). Furthermore, Forsberg & Bartning (2010: 150) showed that the increase was not always statistically significant, something they believed might be due to the low number of essays per level and also the fact that the texts are often fairly short.

Unidiomaticity in learner languages has often (and for a long time) been linked to MWEs (cf. Pawley & Syder 1983) probably due to the fact that there is such a multitude of complicating factors in relation to MWEs. De Cock et al. (2014: 78) claim that the problems with MWEs concern: (1) the extent to which they are used, (2) the MWEs that are used, and (3) how they are used.

How MWEs are used by L1 and L2 speakers have been important issues in recent research within SLA and within learner corpus research, but often from a fairly open perspective on collocations which focuses on how words are used

together with other words based on statistical measures such as MI and log likelihood. However, as stressed by Forsberg & Bartning (2010: 148) these measures require quite large datasets, which is a prerequisite not met by our dataset. We have therefore opted to focus on a knowledge-based approach.

2.2 MWEs and language teaching

Nesselhauf (2003: 223) concluded that MWEs should be seen as “an important part” of L2 teaching, in particular at advanced levels, and that the difficulties which learners experience with collocations require more research. In connection with language learning, lists of MWEs are very useful. There are lexical resources of this kind, but for languages such as Swedish it is hard to find materials with indications of proficiency levels. Furthermore, materials where levels have explicit and transparent information about their grounding in empirical data are quite rare, open access to receptive and productive data being even less common.

One possibility is to use corpora, and several studies have shown that corpora can be useful both to introduce MWEs and to work with *noticing* (Schmidt 2012) strategies (cf. Meunier 2012). Nevertheless, teachers and learners rarely have access to information about how the MWEs occur in learner language or even in data aimed at learners. This is a shame since access to MWEs in data opens possibilities for working with noticing as well as contextualising the usage (see e.g. Boers et al. (2006) who studied how MWEs can be taught with the help of noticing).

Online lexicographic reference sites with information about the MWEs which can be expected at particular CEFR levels are available for English. The English Profile² (Hawkins & Filipović 2012, Green 2012, Kurtes & Saville 2008) is explicitly based on learner data, but does not provide access to frequencies of use or more than the odd example of use in the entry itself. The English Vocabulary Profile (Capel 2012, 2015) makes it possible to select phrases, phrasal verbs or idioms, all of which contain some MWEs. However, there is currently no possibility to select MWEs as a superordinate category. The English Grammar Profile (O’Keeffe & Mark 2017) enables the selection of e.g. phrases/exclamations, expressions with *be*, or items which have been subcategorised as “phrasal”. However, there is no category for MWEs in general. Similarly, Pearson’s Global Scale of English (GSE)³ provides lists of MWEs with proficiency level information, but no frequencies and no access to empirical data which the reference has been based on. The user can choose phrasal verbs and/or phrases. The category ‘phrases’ in

²<https://www.englishprofile.org/>

³<https://www.pearson.com/languages/why-pearson/the-global-scale-of-english.html>

Pearson's GSE includes phatic communication, asking about prices, introducing yourself, and idiomatic expressions and is therefore broader than MWEs, since it deals more with communicative phrases. The Swedish L2 lexical profile, which we are introducing here, provides more information about MWEs. Furthermore, it not only includes frequencies from both receptive and productive empirical data, it also provides access to the data.

2.3 Assigning proficiency levels to lexical items

Even though CEFR focuses on communicative competence, the CEFR documentation (Council of Europe 2001, 2020) still indicates that we should try to associate lexicon items to CEFR proficiency levels. Previous work on assigning levels to words and MWEs based on corpora have mainly focused on coursebook corpora (Gala et al. 2013, 2014, François et al. 2014, 2016, Dürlich & François 2018, Tack et al. 2018), although some works also used learner corpora (Volodina et al. 2016b, Alfter et al. 2016). It has generally been found that a simple method of assigning levels, i.e., using the first level at which an expression occurs, performs better, or at least equally well to more sophisticated methods (Gala et al. 2014, Alfter 2021). However, as the majority of works have focused on coursebooks, it should be noted that other methods of assigning levels, such as threshold approaches, may be more suitable for learner language (Alfter 2021, Yamaguchi et al. 2022). Furthermore, frequency based approaches such as the above may not be well suited to assigning levels to MWEs, as these expressions tend to be less frequent.

2.4 MWEs, compositionality and transparency

Some MWEs such as idioms have often been discussed as examples which go against the compositionality principle in language. However, compositionality is easily confused with transparency, and there is a need to investigate its relation to other features of the MWEs and their constituents (Schulte im Walde 2024 [this volume]). Research on idioms have included debates regarding how semantically analysable idioms are (Cieślicka 2015). These discussions have compared idioms like *spill the beans* and *kick the bucket*, the first has been seen as semantically compositional, since we can imagine each word in the expression being a metaphorical rendering of something: *spill* 'tell something' and *beans* 'secrets' (Nunberg et al. 1994, Cieślicka 2015). However, the second expression is not compositional in this way. This means that an idiom can be seen as decomposable or compositional even when it is figurative and non-transparent. Hence, according to Cieślicka (2015) and Nunberg et al. (1994: 495) decomposability (cf. also compositionality) is not the same as transparency.

2.5 Swedish MWEs

As in international research, there is also a multitude of Swedish terms which have been used in relation to MWEs. Sometimes the Swedish terms are very similar to the English terms, but they do not necessarily mean exactly the same. We aim to make it possible to relate our work to both international and Swedish terminology, which is why we sometimes give both an English and a Swedish term for the sake of clarity. Swedish terms will be preceded by (sv) just as Swedish examples.

Lexicalised phrases, (sv) *lexikaliserade fraser* (lit. ‘lexicalised phrases’), are discussed by Anward & Linell (1976) in a more restrictive sense than the one presented in Sag et al. (2002) and which we have adopted in our study. They focused on a type of lexicalised phrases which have connective prosody where the main stress is on the right-hand part of the expression. They exemplify this with (sv) *en varm korv* (lit. ‘a hot sausage’) ‘a hot sausage’ as opposed to the compound (sv) *en varmkorv* (lit. ‘a hot-sausage’) ‘a hot dog’. Furthermore, according to them, lexicalised phrases can be inflected and syntactically modified internally through separate lexemes. This is sometimes possible in the non-contiguous lexicalised phrases in the Saldo lexicon (Borin et al. 2013) and in our taxonomy, but not always since this is also related to the compositionality and transparency of the expression.

Anward & Linell (1976: 80–81) further divided lexicalised phrases into rather specific subcategories such as premodified noun phrases (NP), e.g. (sv) *Vita huset* (lit. ‘the white house’) ‘The White House’; definite NP with a preposed epithet, e.g. (sv) *profeten Jesaja* (lit. ‘the prophet Jesaja’) ‘Isaiah the Prophet’; definite NP with a postmodifier, e.g. (sv) *Gustav III* (lit. ‘Gustav the third’) ‘Gustav the third’ or (sv) *mannen på gatan* (lit. ‘the man on the street’) ‘common man’; adjectival phrases with prepositional phrasal modifiers, e.g. (sv) *ont i halsen* (lit. ‘sore in the throat’) ‘a sore throat’ etc.

In Swedish linguistics, MWEs ((sv) *flerordsenheter* (lit. ‘multiword-units’) see e.g. Prentice & Sköldbberg 2013) are primarily treated as specific subcategories: e.g. particle verbs ((sv) *partikelverb*);⁴ reflexive verbs ((sv) *reflexiva verb*); idioms ((sv) *idiom*); proverbs ((sv) *ordspråk*) and lexicalised compounds ((sv) *lexikaliserade sammansättningar*). Support verb constructions have also been studied and are referred to as (sv) *funktionsverbförbindelse* (lit. ‘function verb relation’) in the Swedish Academy Grammar (SAG, Teleman et al. 1999); e.g. (sv) *falla i glömska* (lit. ‘to fall into oblivion’) ‘to be forgotten’. Even though the English term *support verb* is very different from the Swedish term, we believe that the meaning is sufficiently close to allow us to use this term.

⁴We call these “particle verb” in English to reflect the Swedish terminology even though they are similar to phrasal verbs.

The term *collocation* ((sv) *kollokation*) is also used in Swedish. Prentice & Sköldb-berg (2013) define collocations as words with a strong association between them and among the examples one can find e.g. (sv) *fatta beslut* (lit. ‘to grab (a) decision’) ‘to come to a decision’ which we would classify as a support verb construction. Exactly what makes something a “strong association” is not clear, but as international literature has often seen “collocations” as a statistical association we believe *MWE*, and the Swedish term *flerordsenhet*, are better terms to use in the context of our work.

An additional complication both in learning and automatically finding and annotating Swedish MWEs is that there are some MWEs which show variation regarding whether they are written as a MWE, or as a single word. For instance, there are adverbs, which are highly lexicalised but which are often written as separate words, e.g. (sv) *i dag* (lit. ‘in day’) ‘today’, (sv) *i går*⁵ (lit. ‘in yesterday’) ‘yesterday’, *över huvud taget* (lit. ‘over head taken’) ‘at all’. The official recommendation for many of these has been to write them apart, but there has been a fair amount of variation, and lately the recommendations have become more relaxed and primarily emphasise consistency (cf. Karlsson 2017, Svenska Akademien 2015). In our manual annotation we currently only annotate the multiword instances of these words and it is only those that appear in the listings in the MWE part of the Swedish L2 Profiles. However, any analysis of these types of MWEs should also consider the single-word variants and it would be good if future work could also include them in the profile next to the multiword instances.

3 Swedish MWE taxonomy

It is important to explore (1) whether some MWEs appear to be easier to learn and (2) which MWEs or MWE types tend to be learnt only at more advanced proficiency levels. Individual MWEs are likely to be highly linked to certain topics. However, since there are many different kinds of MWEs it is interesting to see if learning patterns can be found if we look at how the MWE types occur in both coursebooks for L2 learners and in texts which learners produce, rather than looking at individual MWEs. If so, types could be taken into account more both in teaching and in assessment. In this section we present how we have designed our MWE taxonomy based on previous international research on MWEs, research on Swedish MWEs and in relation to the automatic annotation pipeline we use.

⁵The word *går* is only used in this expression and in the noun *gårdagen* ‘yesterday’ in present day Swedish.

Erman & Warren (2000) distinguished MWEs (formulaic sequences) into lexical, grammatical and discursive ones (prefabs) and applying this taxonomy Forsberg (2008) and Lewis (2008) both found that lexical MWEs were most problematic to L2 learners (L2 French and L2 English respectively) (Forsberg & Bartning 2010). Our taxonomy has a similar division but it is more detailed (see Figure 1). We have two ambitions with our taxonomy:

1. A taxonomy that supports L2 Swedish research, teaching, and learning. It should be connected to what learners might find easy or difficult in learning L2 Swedish. For instance, particle verbs and reflexive verbs can be challenging for learners (cf. Enström 1990, Ekberg 1999).
2. A taxonomy that is computationally useful. While MWEs in this work were automatically identified, our taxonomy could further enhance automatic MWE recognition, which in turn could impact downstream tasks such as parsing efficiency positively.

We want to be able to start from the output of the annotation pipeline. The Sparv-pipeline (Borin et al. 2016) which we use (see Section 4.2 for more details on Sparv) is knowledge-based and depends upon entries currently in the Saldo lexicon (Borin et al. 2013). As part of the lemmatisation, MWEs in texts are identified through Saldo. This means that if the MWE does not have an entry in Saldo it will not be recognised, and if something is not seen as part of a MWE in Saldo it will not be part of the MWE in the list of MWEs which we work with. The latter is the case with certain prepositions since they can either be seen as part of the MWE or as part of the valency of the MWE, cf. (sv) *ha ont (i)* (lit. ‘have ache (in X)’) ‘have a (X) ache’, or (sv) *ta reda (på något)* (lit. ‘take control/organisation (on something)’) ‘find out (something)’.

There are several potential problems with identifying MWEs based on Saldo lookup for work on L2 Swedish:

1. The MWE annotation might not be reliable. There was no previous evaluation of how reliable the Sparv-pipeline is at identifying MWEs, that is, whether it produces too many false positives (overgenerating) or false negatives (undergenerating). We have therefore performed an annotation check as presented in Volodina et al. (2022b) and will summarise and discuss this in Section 5.1 with regards to MWEs.
2. The annotation pipeline may not be reliable on L2 production. L2 production does not necessarily conform to the standard variety of the target

language. This means that the recognition of MWEs is likely to be more complicated, since the pipeline has been trained on fairly normlike texts written (primarily) by L1 speakers. This is also something which we studied in Volodina et al. (2022b) and which is summarised in Section 5.1.

3. The lexicon might not contain the MWEs which are used. Sparv lemmatisation is based on the Saldo lexicon which is far from exhaustive when it comes to MWEs. Borin (2021) claims that MWEs make up 6% of the Saldo lexicon. For comparison, Sag et al. (2002) cite that 41% of WordNet consists of MWE entries (but WordNet entries only include senses for lexical word classes, whereas Saldo also includes senses for grammatical word classes and this is likely to affect the percentage of MWEs). As seen in Section 2, there have been claims that the number of MWEs are equal to the number of single-item entries (Jackendoff 1997), or that there may be ten MWEs to every single item (Mel'čuk 1998). Thus, it is relatively safe to assume that a fair share of Swedish MWEs are not listed in Saldo and would therefore be missed during the automatic linguistic annotation.
4. Saldo does not include 'strong collocates' (i.e. institutionalised phrases). These are also an important "near-phraseological" knowledge for L2 learners and have frequently been looked at in studies of formulaic language in SLA (cf. Section 2). Hence future research needs to find ways to add less lexicalised MWEs to a lexicographic resource aimed at language learners.

In Section 5.1 we show that (1) and (2) are not really an issue and in fact the same checks indicate that (3) also is not a large problem for our data since most MWEs were annotated. We will however have to leave (4) to future research.

In our taxonomy we have tried to take into account previous research on MWEs both regarding second language acquisition and the Swedish language. Our aim is that the categories should be easily relatable to both, and possible to justify formally in such a way that they could facilitate later computational use of the taxonomy.

We focus on conventionalised collocations but we have decided against using the term *collocation*. This is partly because it is often associated with the statistical method of identifying MWEs based e.g. on n-grams or less lexicalised phrases as discussed above. However, *lexicalised phrases* in our sense is sometimes the same as a *collocation* according to others; for instance, Cowie (1994)'s and Nesselhauf (2003)'s use of collocation relies on there being an "arbitrary restriction on substitutability" (Nesselhauf 2003: 225) which is similar to our idea of lexicalised phrases.

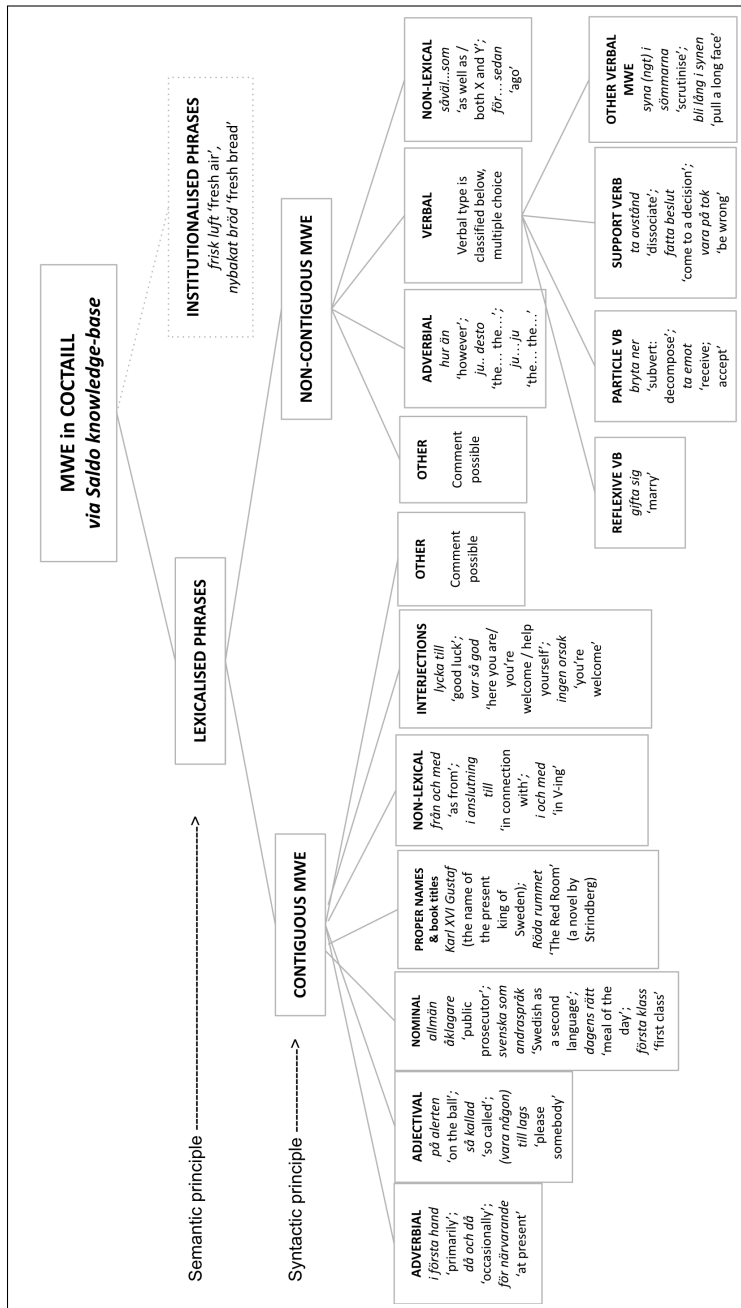


Figure 1: Our MWE taxonomy, below each category there are some Swedish examples to help the annotators remember the definition.

In our taxonomy (Figure 1), we divide MWEs into *lexicalised phrases* (cf. Sag et al. (2002)) and *institutionalised phrases* (cf. “strong collocates” or Cowie’s “free combinations”). The latter are not currently included in the Sparv pipeline and are therefore not included in our current work. Lexicalised phrases are further divided into *contiguous* and *non-contiguous MWEs*⁶ according to syntactic principles and later subcategorised according to word class where the syntactic function associated with different word classes is of particular importance (cf. Leseva et al. 2024 [this volume]). The contiguous MWEs are therefore divided into: adverbial MWEs, adjectival MWEs, nominal MWEs, non-lexical MWEs, but also proper names where we include book titles. One might not need to learn book titles, but we classify any which occur since they are a type of MWEs and some of them are such that they can be expected to be part of the common knowledge of Swedish speakers, e.g. *Röda rummet* ‘The Red Room’, a famous novel by the Swedish nineteenth century author August Strindberg. Finally, we also include a contiguous category of interjections since e.g. greetings are common in learner language and often consist of lexicalised MWEs.

Among the non-contiguous MWEs we find adverbial MWEs, verbal MWEs and non-lexical MWEs. The verbal category contains several sub-categories which are often mentioned in both teaching and research. Therefore we annotate if the verbal MWE is: a reflexive verb e.g. (sv) *gifta sig* (lit. ‘to marry oneself’) ‘to marry’, *lära sig* (lit. ‘to learn oneself’) ‘to learn’, a particle verb e.g. *bryta ner* (lit. ‘to break down’) ‘to subvert, to decompose’, *ta emot* (lit. ‘to take against’) ‘to accept’, a support verb construction e.g. *ta avstånd* (lit. ‘to take distance’) ‘to dissociate’, *fatta beslut* (lit. ‘to grab (a) decision’) ‘to come to a decision’, *vara på tok* (lit. ‘to be on mistake’) ‘to be wrong’, or some other type of verbal MWE. This final category includes more idiomatic expressions such as (sv) *syna något i sömmarna* (lit. ‘inspect something in the seams’) ‘check something carefully’ and (sv) *bli lång i synen* (lit. ‘become long in the sight’) ‘pull a long face’. In addition, both contiguous and non-contiguous MWEs can be classified as “other” and a comment can be added since some items could be difficult to annotate into these categories and it is important to be able to come back to any such items at a later point.

During initial annotation we included the categories of *idiom* and *fixed expression* which were removed before the final annotation. They proved to be too problematic to define in such a way that they did not overlap with each other or

⁶Cf. *continuous* and *non-continuous MWEs* in this volume. Since we are presenting our taxonomy here, we need to use the terminology we have chosen there. The terminology we chose for our taxonomy is also what is used in our online resource.

with other categories such as interjections. What one annotator saw as a fixed expression was sometimes seen as an interjection by the other, e.g. for greetings.

There are also some other distinctions which can sometimes be a bit problematic, e.g. (sv) *nybakat bröd* (lit. ‘newly baked bread’) ‘fresh bread’ is categorised as an institutionalised phrase, while (sv) *dagligt bröd* (lit. ‘daily bread’) ‘daily income’ is categorised as a nominal MWE. In a religious context, which is probably the most common context for the expression, this could also be considered an “institutionalised phrase”, but, this is not really the sense of institutionalised phrases in our taxonomy.

We discussed making further divisions according to transparency and compositionality, and experimented with annotating compositionality on a scale from 0–100. However, rating the compositionality (or transparency, see further Section 4.2) proved very difficult to do in a systematic way for the annotators, and their annotations indicated that they might have interpreted the concept of compositionality differently.

4 Materials and methods

In this section we introduce the corpora which we have used for this study (Section 4.1), and how we have automatically identified and manually annotated MWEs in them (Section 4.2). We then briefly describe how we have checked the automatic MWE annotation (Section 4.3). In Section 4.4 we explain how we have linked the MWEs to levels and also summarise a couple of studies where we have studied whether crowdsourcing could be used to link lexical items to proficiency levels. Since this chapter also demonstrates how the annotation can be accessed and used for further analysis of MWEs, we also introduce the graphical user interface of the Swedish L2 Profiles here (Section 4.5).

4.1 Corpora and Sen*Lex

We study the development of L2 Swedish based on two Swedish corpora: Coctail (Corpus of CEFR-based Textbooks as Input for Learner Levels’ modelling, Volodina et al. 2014), a corpus of coursebooks used in teaching L2 Swedish to adults, and the SweLL-pilot corpus (Swedish Learner Language Pilot corpus, Volodina et al. 2016a), a corpus of L2 Swedish essays. Coctail is used as a representation of receptive proficiency, however it can also be used as a proxy for common input at the different CEFR levels. The SweLL pilot is used to study the productive proficiency at different levels.

The corpora have been processed automatically using the Sparv pipeline (Borin et al. 2016), including tokenisation, lemmatisation, part-of-speech (POS)-tagging, dependency parsing, and word sense disambiguation. The pipeline also identifies MWEs during the process of lemmatisation based on a knowledge-based method that makes use of the Saldo lexicon (Borin et al. 2013). We use this automatic MWE annotation as a basis for our further manual annotation, i.e. only MWEs identified in this process are additionally annotated. We have also evaluated the success of the automatic annotation on a variety of texts which we use: coursebook texts and learner texts from different proficiency levels (see Section 4.3 and for results see Section 5.1).

The corpora have previously been used to derive two lexical resources aimed at language learners: (a) a CEFR-graded resource for Swedish as a second language, SVALex “SVenska som Andraspråk Lexikal resurs” (lit. ‘SWedish as a Second language Lexical resource’) (François et al. 2016) which is based on Coctail and which shows the expected receptive knowledge, and, (b) SweLLex (Swedish Learner language Lexicon, Volodina et al. 2016b) based on the SweLL-pilot corpus and which targets productive knowledge. In these lexical resources you can find the lemgrams (i.e. lemma + part of speech) of the words which occur in the corpora, but homographs are not separated if they have the same part of speech and the same inflectional paradigm. Hence, (sv) *val* ‘election’ and (sv) *val* ‘whale’ are different lemgrams since their inflectional paradigms are different: (1) *ett val*, *flera val* ‘one election, several elections’, (2) *en val*, *flera valar* ‘one whale, several whales’. But (sv) *gå* which can mean several different things including ‘walk’, ‘go’ or ‘be possible’ cannot be distinguished based only on lemgrams, therefore, word sense disambiguation is needed.

We recreated the lists with automatic word sense disambiguation, resulting in a list where each item includes lemma + POS + sense. The resulting lists are called SenSVALex and SenSweLLex, but are usually treated as one and referred to as *Sen*Lex*, cf. Alfter (2021: 31–32). *Sen*Lex* includes a total of 16 324 items (excluding some problematic cases but including MWEs). Any of these items which has been annotated as MWE are categorised manually according to MWE types (cf. Section 4.2).

4.2 Automatic and manual MWE annotation

Bearing in mind the limitations of Saldo as a knowledge base for the identification of MWEs (see Section 3), we nonetheless argue that the Saldo-based, i.e., knowledge-based, MWE identification is useful, objective and reliable (cf. the annotation check results in Section 5.1 and Volodina et al. 2022b). We therefore

explore the phraseological dimension of L2 vocabulary starting from the MWEs identified through the Sparv pipeline (Borin et al. 2016) and based on Saldo (Borin et al. 2013). Two annotators⁷ categorised the MWEs further into subcategories (cf. Section 3).

All automatically identified MWEs were added to a database for further manual annotation in Legato (Alfter et al. 2019), where only the types in our taxonomy could be selected. The manual annotation was done according to guidelines.⁸ An initial round of annotation was analysed and resulted in a modified taxonomy as presented in Section 3, as well as clarifications in the guidelines which the annotators made use of during the second and final round of annotation. The first author was available for supervision during annotations.

In the first round of the manual annotation our annotators were asked to indicate compositionality. This was excluded from the second round because it seems that our annotators understood the concept of compositionality differently. Previous research has also shown that compositionality is sometimes confused with transparency and vice versa (Cieślicka 2015, Nunberg et al. 1994). Instead, we decided to ask if contiguous MWEs were (morpho)syntactically modifiable or not in the final annotation.

Compositionality and transparency in relation to MWEs is definitely an area that requires further investigation, and a larger experiment with rankings of transparency and compositionality would be very interesting. However, this can only be done if a better way can be found of either defining or operationalising the concepts. Nevertheless, when there appeared to be a large difference between the ranking from the crowdsourcing experiment and the level of first occurrence in the coursebooks we decided to compare these results to the initial annotations from one of the annotators regarding the “compositionality” of the MWEs. We compared the relative rankings from the crowdsourcing experiment to the compositionality judgements. We did not use both since they seemed to interpret the task or concept differently.

4.3 Annotation check

As part of a more comprehensive annotation check (Volodina et al. 2022b) we also checked the annotation quality of MWEs. In this chapter we will summarise

⁷Both annotators are L1 speakers of Swedish. One has a MA in Scandinavian languages and one has a PhD in the same.

⁸<https://docs.google.com/document/d/1nZOKf-54FEkjIQFnPUmZZRWqib6y7gpCuKQO-XadeqM/edit?usp=sharing>

the parts about MWEs which have previously been published and discuss the results further.

The check was done by letting two annotators⁹ go through the automatic annotation of three texts per level (5 levels A1–C1) from three different sources: (1) the coursebook corpus, Coctail, (2) the original learner corpus, SweLL-pilot, and (3) the same as (2), only normalised which, among other things, standardised the spelling (cf. Rudebeck et al. 2021 for the normalisation guidelines, which we followed to facilitate comparisons).

Each annotator received a spreadsheet with the texts and their annotation in one tab per source (coursebooks, original learner texts, normalised learner essays respectively). Each annotation type was presented in a column of its own next to which a separate column was used for corrections. There was one token per row, which meant that MWEs spread over several rows. Next to the column of lemma, or MWE annotation, there was an extra column for corrections. The columns which should not be changed were locked so that only one or two of the researchers had access to them. The check was done according to a set of guidelines and under the supervision of the first author.

After the check was finished by both annotators a first comparison was run by one of the researchers, comparing the cells in the columns for corrections. Then a more qualitative check was done where the first author checked the changes that had been made to any items in relation to MWEs, i.e. additions or deletions of tokens from MWEs which had been identified, as well as if any MWEs had been noted as having been missed completely (undergeneration) or identified mistakenly (overgeneration) (see Section 5.1). An analysis of the whole check has been published in Volodina et al. (2022b).

4.4 Proficiency level assignment

CEFR levels are assigned based on the appearance of items in L2 Swedish coursebooks, as found in Coctail, and in L2 Swedish learner essays, as found in SweLL-pilot (cf. Section 2.3). Of course there are lexical items which do not occur among the words in the corpora and we therefore wanted to see how those could be linked to CEFR levels, but also how alternative ways of linking levels to MWEs would compare to the levels of lemmagrams which had already been graded based on coursebooks and learner essays. For this reason, we tried to rank MWEs through crowdsourcing by experts and non-experts, as well as by direct labelling by experts (Alfter et al. 2021). These results are not available in the Swedish L2

⁹Both annotators have an MA in Scandinavian languages. One is an L1 speaker of Swedish and the other an advanced L2 speaker.

Profiles but they are partly presented in Section 5.2. The crowdsourcing experiment meant that participants were asked to say which out of four MWEs they judged to be the easiest, and which they judged as the most difficult. We sorted the MWEs into three groups which were presented in separate experiments:

1. Group 1: Interjections, fixed expressions and idioms;
2. Group 2: Verbal MWEs and
3. Group 3: Adverbial, adjectival and non-lexical MWEs.

In parallel with the crowdsourcing experiment, we also asked three L2 Swedish professionals who had good knowledge of CEFR to first go through all of the crowdsourcing tasks, and then perform an explicit ranking assignment in a spreadsheet, where they had to assign CEFR levels from A1 to C2+ to each MWE which was part of the crowdsourcing experiment (Alfter et al. 2021).

Apart from the quantitative analysis in Alfter et al. (2021), we have analysed some of the results qualitatively in a previous publication (Lindström Tiedemann et al. 2022) and some of the main results from the latter are summarised in Section 5.2.1. This analysis was based on the MWEs which had been ranked as easiest and hardest in the different participant groups and included the seven easiest and the seven most difficult items from each type of MWEs and from each participant group. The MWEs were compared qualitatively across groups, and also in relation to the fact that if all groups had picked the same items as the seven easiest there would be 21 in total (7×3 MWE types). The results were also compared to coursebook occurrences, and newspapers and blogs to some extent, as well as to the direct ranking results.

In this chapter we present a further qualitative analysis of some of the results from the crowdsourcing experiment. This will help us gain a better understanding of why some expressions seem to be ranked very differently in the coursebook rankings in comparison to the general implicit rankings by the crowdsourcers. We examine the items in group 1: interjections etc., from the crowdsourcing experiment. This was the group with the best correlations in the crowdsourcing experiment, but still some results were a bit surprising in comparison to the first occurrence in coursebooks, and it would be interesting to see if those have anything in common.

Since we originally picked twelve items per level from the coursebooks, and since the crowdsourcing experiment results in a continuum from 1–60 we make a naive working assumption that 12 steps equal one level, even though this certainly does not have to be the case. We do this just as a way of deciding on a

selection principle for the items to look at more closely. We therefore focus on items that were one level away from the level they were chosen to represent, based on the occurrences in the coursebooks, or more than one level. We check if the same expression occurs in more than one coursebook aimed at the same level, to see if they could be seen as *core items* that several books considered important to include at the same or adjacent levels (Volodina et al. 2022a).

4.5 MWEs in the Swedish L2 Profiles

Based on the manual annotation described in Section 4.2 above, MWEs can now be accessed and filtered in different ways in the lexical profile within the Swedish L2 Profiles. There we provide lists of MWEs which appear in coursebooks for L2 Swedish learners (the Coctail corpus) or in learner essays (the SweLL-pilot corpus) including information about the proficiency levels where they appear in coursebooks (presented as receptive) and learner essays (presented as productive). The information can be filtered according to part-of-speech, MWE type, or CEFR level.

The profile includes absolute and relative frequencies and links to Korp (Borin et al. 2012) at Språkbanken Text where the corpus evidence can be consulted.¹⁰ This is what we use for our analysis of how MWEs occur in the data Section 5.3, and it is openly available to other researchers and teachers who wish to explore the resource.

5 Analysis and results

In this section we first present the results of our annotation check where we wanted to see how well the Sparv-pipeline identifies MWEs in both coursebooks and learner texts (Section 5.1). This includes a discussion of our results in Volodina et al. (2022b). In Section 5.2 we analyse different ways of assessing the proficiency level which should be associated with different MWEs. Finally, we analyse the distribution of MWEs across proficiency levels based on their occurrence in coursebooks and learner essays as presented in the Swedish L2 Profiles (Section 5.3).

5.1 Quality of the automatic annotation of MWEs

We focus on lexicalised MWEs such as verbal MWEs (e.g. particle verbs and reflexive verbs), greetings, multiword prepositions. These might not occur that of-

¹⁰The productive data requires a licence to access the actual texts.

ten in texts and they can be non-contiguous or allow some variation which can complicate their automatic recognition. For this reason we have opted to use a knowledge-based approach to MWE which we presented above (Section 4.2) and which we present an evaluation of here based on Volodina et al. (2022b), and here we also discuss these results further.

The automatic MWE annotation works best on coursebook data which has been written by L1 speakers. It also works better on the normalised L2 data than on the original L2 data, indicating that original L2 data is a bit more problematic as expected. There is no clear correlation between issues in the annotation check and certain proficiency levels in either of the datasets, instead the precision and recall seem to vary idiosyncratically, but we suspect it may be related to genre, topic and task type (Volodina et al. 2022b).

Overall, the MWE annotation works fairly well: 7–8 out of 10 MWEs were correctly annotated as MWEs, 2–3 were missed and some items were labelled as MWEs even though they were not MWEs, or they were only partially recognised (Volodina et al. 2022b: 158). In 45% of the missed cases it turns out that the MWE is also missing in the Saldo lexicon (Volodina et al. 2022b) which, as mentioned above, was one of the weaknesses we expected to see when using a knowledge-based MWE annotation system. Still, since MWEs are mostly well annotated and since the MWEs which are included in Saldo are likely to be the most well-known in Swedish, we believe this is a good result. It is clear that this could be improved by adding more items to Saldo, e.g. based on the items we have found to be missing. Nevertheless, it is clear that some MWEs which are listed in Saldo were also missed in the annotation, and it would be a good idea to look at these instances in more detail in the future to see if the annotation could somehow be improved, and maybe such work could also contribute to our understanding of MWEs.

Checking MWE annotation seems to be cognitively more difficult than checking lemmas, POS etc. (Volodina et al. 2022b). Furthermore, there are several MWEs that include a placeholder and such MWEs are not yet fully part of Saldo even though similar constructions have been studied extensively in relation to the Swedish Constructicon (see e.g. Sköldbberg et al. 2013, Lyngfelt et al. 2018). This is related to the difficulty of annotating such instances automatically and requires further research.

Our assistants who checked the annotation agreed quite well in the MWE check. Krippendorff's alpha (Krippendorff 1980) show inter-annotator agreement at 0.85 for Coctail, 0.74 for original learner essays and 0.89 for the normalised learner essays, the highest value (Volodina et al. 2022b). This could possibly be because the normalised version of the learner essays was checked more closely in time by the two assistants and hence the discussions with the supervising

researcher might have been more similar in relation to this set. The annotation of Coctail and SwELL original had been checked quite a long time before then by assistant 1 and the check had resulted in some discussions about annotation and annotation check practices. This also partly meant that assistant 1 was allowed to go back and change her annotations to some extent and the guidelines were clarified for assistant 2 on these accounts.

Differences in the MWE check show that the assistants partly disagreed on what a MWE is, or differed in how good they were at spotting certain types of MWEs: one recognising grammatical MWEs such as (sv) *trots att* ‘even though’ more easily, and the other recognising complex verbs such as (sv) *få barn* (lit. ‘receive (a) child/children’) ‘have a child/children’ more easily (Volodina et al. 2022b). This could also partly have been affected by the fact that after assistant 1 started checking the data we saw that many MWEs were seen as missing a preposition (cf. Section 3) which led to discussions with Saldo staff who explained that prepositions are usually not seen as part of the MWE, but rather as part of the valency of the MWE and therefore are not listed in Saldo. Assistant 2 received this information before the check and hence could bear it in mind from the start.

5.2 MWE and proficiency levels

The results of the crowdsourcing experiment show that non-experts and experts rank MWEs very similarly, whereas direct ranking seems to be difficult and show less agreement between the annotators and also show some disagreement with the crowdsourced relative rankings by the same expert (Alfter et al. 2021). Unfortunately, since we chose to use relative judgement in the crowdsourcing experiment we have not yet found a way of directly linking items to a particular proficiency level. Instead the results are on a continuum with no indication of precise levels.

The fact that explicit level assignments seems to be more difficult and less consistent is something which we have concluded correlates well with previous studies on assessment of proficiency, although most of those studies have been on data consisting of full texts rather than decontextualised expressions which are bound to be even more difficult to link to an explicit level (Alfter et al. 2021).

5.2.1 The easiest and the most difficult items

In a previous qualitative analysis of the items which were ranked as the easiest seven or the most difficult seven in each of the three sets (i.e. interjections, verbal, adverbial) and for all three groups (i.e. learners, teachers, experts) we found

that the crowd participant groups agreed fairly well, but they agreed more on the easiest words (Lindström Tiedemann et al. 2022). This is hardly surprising since the topics are also more clearly defined in relation to the lower CEFR levels than the higher levels which are more associated with professions and special interests.

There were 28 MWEs in total which appeared among the easiest seven – instead of 21 which would have been the case if experts and non-experts had been in complete agreement (7 x 3 groups of MWEs). Thirteen (46.4%) of the easiest MWEs were among the easiest seven for both expert groups and for the non-experts. Five (17.9%) were among the easiest according to the L2 speakers (non-experts) and one of the expert groups. Hence there was partial agreement for eighteen MWEs (64.3%). The most difficult seven expressions were as many as 35 MWEs (compared to 21 which would have meant total agreement on the seven expressions, but not necessarily on their order). Only nine (25.7%) were the same in all three groups, and an additional seven (20%) showed agreement between the L2 speakers and one of the expert groups. That is, there was only 45.7% partial agreement among the most difficult items.

Comparing these results to the rankings based on the coursebooks showed that 18 (85.7%) of the easiest MWEs picked by the L2 speakers appeared at A1 level in coursebooks, somewhat less of the items picked by the experts: 76.2% of the L2 teachers' and 71.4% of the CEFR experts' (Lindström Tiedemann et al. 2022). Hence the L2 speakers' relative judgement was in a sense more in line with coursebook rankings for the easiest expressions. The most difficult expressions were harder to compare to coursebook occurrences. L2 speakers had nine (43%) MWEs from C1 among their most difficult seven, L2 teachers eleven (52%) and CEFR experts seven (33%).

Some of the MWEs were ranked as quite easy in comparison to the coursebook-based levels and sometimes it seemed as though this could be because the expressions were commonplace everyday expressions (Lindström Tiedemann et al. 2022). Prentice & Sköldberg (2013) claim that MWEs can be more common in informal genres. Some of the MWEs which were ranked as easy clearly do appear more in blog corpora than in newspaper corpora (Lindström Tiedemann et al. 2022) which could be a sign of this, but this should be investigated further. Other expressions which were ranked as easy were clearly more international expressions such as (*sv*) *logga in* 'to log in', which was seen as relatively easy by all three groups, but which only occurred in coursebook texts at C1 level, except for appearance in an exercise at A2 level. In addition, when ranking MWEs explicitly to levels, the three CEFR experts ranked this item as A1 or A2. This shows that it might be important to include all vocabulary from coursebooks, also from

exercises, and not only from readings texts, since some lexemes might only be used in exercises.

5.2.2 Comparing relative ranking to coursebook occurrences

Since the crowdsourced rankings are relative, they cannot easily be compared to CEFR levels. Still, it is interesting to try to do so by focusing on items which appear to have been ranked quite differently from the level which they were picked for. Since we picked 12 items per level we here focus on items which in the relative ranking ended up one level (eleven steps on a continuum from 1–60) or more, from their level based on their first occurrence in the coursebooks. Some of these MWEs do not even occur at adjacent levels in different coursebooks. In many cases where the discrepancy between the ‘levels’ was large we only have evidence from one book at the level of first occurrence. There are 20 cases (33%) where the ranking has a discrepancy of eleven or more steps on the scale from 1–60. Out of these 17 (85%) have a first occurrence based on one single book, and hence cannot be considered core items at that level, and this could be the reason why they have been ranked quite differently in the crowdsourcing experiment. In six (30%) cases we have examples from other levels, of which two (10%) include the level that was estimated based on the average implicit ranking by the crowd if we assume that the first 12 items on the continuum of 60 items equal A1, the second 12 A2, etc.¹¹

5.2.3 Comparing relative ranking to compositionality or transparency

Fourteen of the twenty cases which appeared to be ranked quite differently to the level which they were picked to represent based on coursebook occurrences had non-compositionality scores which were $\geq 50\%$ in the manual annotation. In fact, for as many as 13 it was $\geq 69\%$; and for nine it is as high as 98–100%. Therefore it is possible that these items were ranked as rather difficult in the crowdsourcing experiment due to their non-compositionality or opacity, since this meant that their meaning could not be guessed based on the combination of the constituent words. High non-compositionality meant that the items were more difficult than expected based on the first occurrence in coursebooks (if we estimate that 1–12 on the continuum from 1–60 should be equivalent to A1 and 48–60 to C1). The only two items (out of the 20) which were seen as opaque and which still received a

¹¹This was used as a naive principle to relate the continuum to levels as a means to clarify the differences and facilitate the analysis. However it is of course possible that all items in the relative ranking are seen as equivalent to A1 or C1 items by the participants.

‘lower’ implicit ranking than the coursebook projections were (sv) *spetsa öronen* (lit. ‘to sharpen the ears’) ‘prick up one’s ears’ and (sv) *många järn i elden* (lit. ‘(to have) many irons in the fire’) meaning ‘(to be) busy’. They were both classified as one level lower in the implicit ranking (if we assume that the order should be equivalent to 12 items per level).

Items that were ranked as easier by the expert and non-expert crowds were often classified as quite compositional (30–69, average 44). And conversely items which were classified as more difficult were seen as relatively non-compositional by the same annotator (74–100, average 88). To conclude, it could be that receptive levels of MWEs can often be correlated to compositionality or transparency, however it is unclear which it was that annotator focused on.

5.2.4 Comparing relative ranking to L1 reference corpora

There was one item in group 1 that the non-expert crowd and the expert crowd agreed was most/second most difficult, namely (sv) *på pin kiv* ‘just to tease’. This item was classified as C1 based on the coursebook texts. It appears three times, but only in one of the books on C1 level.

På pin kiv is a rare expression and 100% non-compositional or opaque according to our annotation. It occurs only rarely 0.1/1 million tokens (25 actual occurrences) in the Swedish newspaper corpus *Göteborgsposten (GP) 1994–2013*, and the same, 0.1/1 million (2 actual occurrences) in the Finland-Swedish newspaper corpus *Hufvudstadsbladet*. It does not occur at all in Finland-Swedish blogs (corpus *Bloggtexter*), but it does occur at the same relative frequency as in the newspaper corpora, 0.1 (35 times), in Swedish-Swedish blogs (corpus *Bloggmix 2006–2013*).¹²

The constituent (sv) *pin* is a rare adjective which only appears in this MWE and hence works as a good immediate idiom key. There are homonyms: (sv) *pin* (noun) which is only used in the Swedish proverb *vill man vara fin får man lida pin* (lit. ‘he/she who wants to look nice must suffer pain’) ‘good looks hurt’ where *pin* is interpreted as a form of pain, and (sv) *pin* (adverb) meaning ‘to the highest level’ used in the compound, e.g. (sv) *pinfärsk. Kiv*, similarly, is quite rare and its only meaning in one of the most authoritative Swedish dictionaries (Svenska Akademien 2021) is “minor disagreement”. Based on both the opaqueness of the constituents of the MWE and the rare occurrences in L1 corpora it is quite understandable that this was ranked as among the most difficult items for L2 learners by experts and non-experts alike.

¹²All corpora were accessed through Korp at <https://spraakbanken.gu.se/korp/>

Interestingly, although not that surprising, the two easiest items in group 1 were also ranked the same in all three crowd participant groups: (sv) *god morgon* ‘good morning’ and (sv) *god natt* ‘good night’. Furthermore, these items were classified as 45–46% non-compositional and they are fairly transparent and also very common greetings. Greetings are among the lemmas that are usually taught explicitly at the beginning of language courses.

5.3 Frequency and MWE types

MWEs make up 4.9–9.4% of the sense-based lemgrams from the receptive corpus and 1.4–9.4% from the productive corpus (cf. Figure 2). The percentage is higher in the receptive data all the way up to C1 level. However, the difference decreases steadily. The percentage is thus much lower than the c. 50% which Jackendoff (1997) estimated that MWEs should make up of our mental lexicon. However, our estimates are higher than Borin (2021)’s for all of the Saldo lexicon, which he claimed consisted of 6% MWEs.

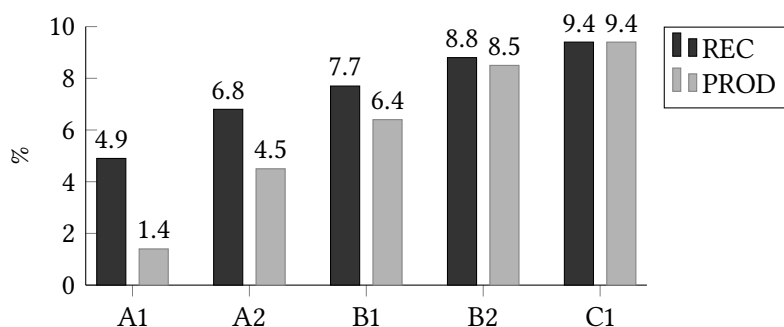


Figure 2: The percentage of MWEs on our sense-based list of lemgrams from Coctail (REC) and SweLL-pilot (PROD)

The fact that data produced by L2 learners reach the same percentage of MWEs in the texts at C1 level as the percentage used in coursebooks at the same level seems very encouraging. It would be good to compare this with other corpora, but to do so in a reasonable way we would need to extract a list of types consisting of lemgram + sense in the same way as here and that will have to be left for future work. In addition, we should try to find a way to include MWEs which are partly non-normlike in the L2 data and which therefore have not been annotated by the pipeline, e.g. due to a mistaken form or non-normlike word order or even a word that is the wrong word in the context but synonymous. To study MWE acquisition we need to find ways to compare such occurrences to the rest; but, it

seems they can only be captured manually unless they are consistent ‘mistakes’ which many learners tend to make.

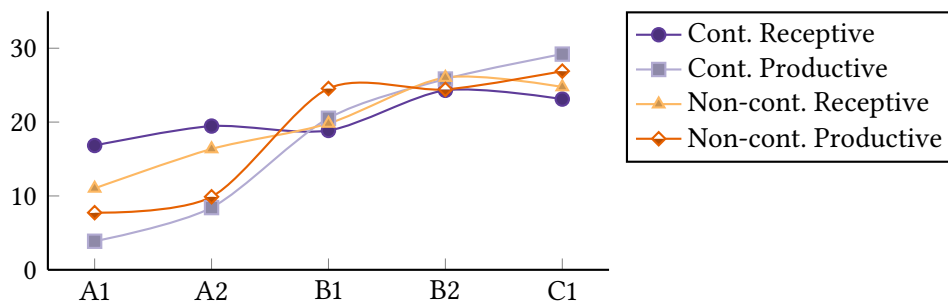


Figure 3: Development of contiguous and non-contiguous MWEs over CEFR proficiency levels in both coursebook data (receptive) and learner data (productive). The frequencies are based on the number of types in each category, not the frequency of occurrence of each type.

Contiguous and non-contiguous MWEs show similar development over the CEFR levels as shown in Figure 3. There is a clear increase in both broad types of MWEs, and from B1 level contiguous MWEs have very similar levels in receptive and productive data. Surprisingly, there are more non-contiguous MWE lemmas in productive learner data from B1 than in the coursebook data. But from B2 the levels are similar in production too. The higher B1 level seems likely to be due to a task effect (cf. Caines & Buttery 2017). Of course we know that in the productive data we are only catching items which have been correctly spelled and might therefore miss some instances which are reasonably normlike and which are attempts to produce MWEs which learners are starting to grasp. It is also possible that some MWEs have been used in the wrong context or in an unidiomatic way. For instance, the preposition might be wrong, since prepositions are not usually included in the MWE in Saldo. This means that a non-standard preposition will not cause the MWE to go unnoticed by the system. This makes it even more interesting that at B1 level there were more non-contiguous MWE lemmas in the learner data than in coursebooks. Including less, in the MWE itself, proves to be a possible advantage when working with learner texts. If MWEs can be identified even when they are not used in a normlike manner, that facilitates a closer analysis of how normlike the usage is.

Zooming in on the subcategories among the contiguous MWEs (Figures 4 & 5), adverbial MWEs stand out at all levels in both the receptive and the productive data. It would be interesting to compare this to other genres. It is also striking that at A1, in the productive data, the only contiguous MWE type is contiguous

adverbial MWEs, but at A2 there are already several others. The only subcategory among the contiguous MWEs which is still missing in the productive A2 data is proper nouns and this could possibly be related to pseudonymisation of the data, although it seems unlikely in this case since the MWE proper nouns that are included in Saldo are generally famous people, companies, organisations etc. This means that it is more likely to be the tasks at A2 that are restricting the use of MWE proper nouns. The category is still missing at B1 and is quite rare at B2 and very rare at C1.

Interjections are quite rare in productive data and there is a decrease in the use of MWE interjections in the receptive data. Still, it is interesting that there are MWE interjections at all levels in the coursebooks, even at C1 level. It would be interesting to have a closer look at the interjections that appear and in which text types they are used in the coursebooks. One would expect that they would only be used in dialogues, but considering the fairly high number of lemmas at more advanced levels when dialogues are more rare this requires further analysis. The fact that they are rare in learner language is most likely due to the tasks.

All non-contiguous MWEs in our data are verbal (Figure 6). The most common subcategory by far is particle verbs in both receptive and productive data. The frequency clearly increases from A1 to B2 in the receptive data and then drops a bit at C1. In the productive data there is a drop at A2. Otherwise there is a steady increase, so it seems likely that the drop at A2 is task-related or a sign that they are starting to really produce MWEs which they know and not whole-phrase constructions. In the productive data there are no instances of the categories *other* or *reflexive verbs* at A1 and interestingly there are no support verb constructions at C1, this may well also be task related.

Reflexive verbs start to appear at A2 in the learner data and show a clear increase at B1 and then remain quite stable. After this the levels are similar to the coursebook data possibly indicating that a ceiling has been reached. The number of lemmas are then slightly higher than in the coursebooks, but this is probably due to the topics covered in the respective texts. Reflexive verbs are quite common in different languages. However, the verbs which are reflexive differ even between closely related languages (Enström 1990: 41) (e.g. (sv) *lära sig* (lit. 'teach oneself') 'to learn'). In fact, they differ even between varieties of the same language, e.g. (sv) *ändra sig* (lit. 'change oneself') 'change one's mind' is not necessarily reflexive in Swedish as spoken in Finland (cf. af Hällström & Reuter 2008). Björklund (2007) similarly mentions differences in the usage of e.g. (sv) *köpa sig* (lit. 'to buy oneself') 'to buy' and (sv) *köpa* 'to buy' in Sweden and Finland, which however could be due to the small size of her data.

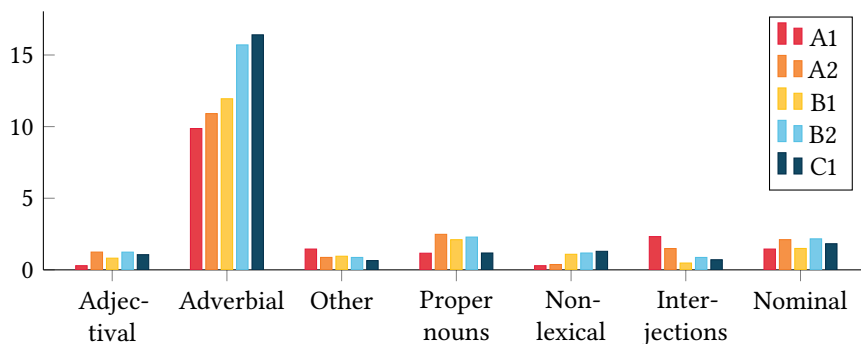


Figure 4: Contiguous MWE types in coursebook data (receptive) over CEFR levels per 10 000 tokens, i.e. the frequency of occurrence (cf. tokens) of the different types (cf. lemmas) is not considered here.

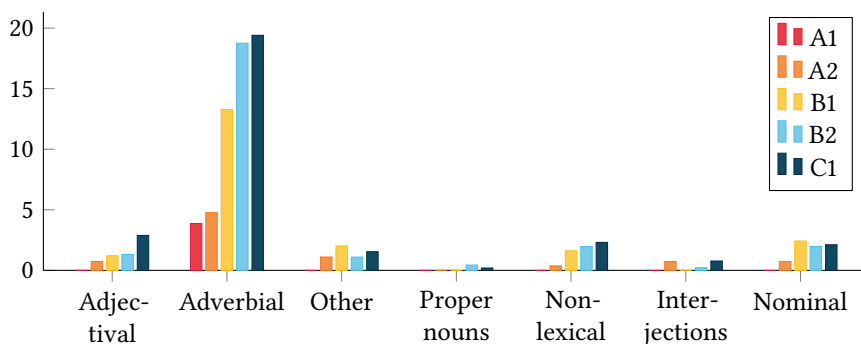


Figure 5: Contiguous MWE types in learner data (productive) over CEFR levels per 10 000 tokens, i.e. the frequency of occurrence (cf. tokens) of the different types (cf. lemmas) is not considered here.

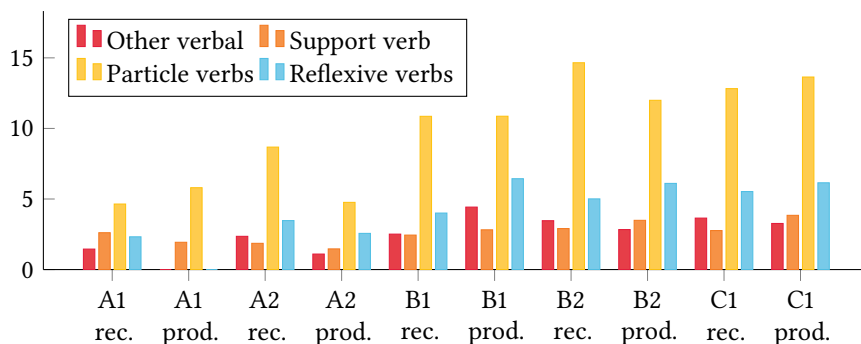


Figure 6: Non-contiguous MWE types in relation to 10 000 tokens, i.e. the frequency of occurrence (tokens) of the different types (cf. lemmas) is not considered here.

It is quite possible that there may be interference from previously known languages which cause difficulty in learning which verbs are reflexive. It is likely that this is why it is quite common that the reflexive pronoun is left out. Furthermore, in relation to usage-based theories of language acquisition how easy it is to learn that a verb is reflexive or not could depend on which variety of the target language you have been in contact with more, if there is indeed variation with regards to whether the verb is reflexive in different varieties. Nevertheless, previous research has shown that there were not that many mistakes in connection with Swedish reflexive verbs, but they seemed to be underused both on the type and the token level in a comparison between L1 and advanced L2 writers (Enström 1990: 93–94).

The fact that Enström (1990) found that there were not many actual mistakes in comparison with the standard regarding how the reflexive verbs were used means that a quantitative comparison based on the reflexive MWEs is likely to give quite a correct image of the use in both coursebooks and learner texts. It is particularly interesting to see that the frequency of reflexive verbs at B1–C1 in our data is higher in the learner essays than in coursebooks. Quite possibly this is related to the essay topics, but it should be investigated further.

6 Conclusions and future work

The Swedish L2 Profiles provides access to Swedish MWEs sorted by CEFR levels and with a possibility of comparing the usage in receptive and productive data by frequency as well as in context. This makes it a versatile resource e.g. for researchers and teachers, and it is clearly different from the English Profile and the Pearson Toolkit. Not only do MWEs occur in one place with a possibility to sort by types, but it also gives open access to frequencies and all of the empirical corpus evidence and presents *productive* and *receptive* overviews side by side.

In this chapter we have summarised how we have ascertained that the knowledge-based automatic MWE annotation works well enough for future studies of L2 Swedish. We have also presented our MWE taxonomy where we have tried to find an appropriate way of building on previous research and making it useful also in connection to automatic annotation. The first round of manual annotation showed that some MWE types can be difficult to keep apart, but after adjusting our taxonomy the two annotators agreed quite well.

Our attempts to find new ways of linking MWEs to receptive proficiency levels proved to work well only as relative measurements, not in relation to discrete levels (Alfter et al. 2021). The best agreement seems to be with regards to ranking

easy MWEs rather than difficult ones. This seems to have support in the CEFR documentation (Council of Europe 2001), since the beginner levels have a very clear focus on certain topics and themes, whereas advanced levels are more varied because they relate to different areas of expertise (cf. Lindström Tiedemann et al. 2022).

In this chapter, we have extended our analysis of the crowdsourcing results through a qualitative analysis of items which were ranked differently than their coursebook first level of occurrence. The results indicate a possible tendency for items which occur in only one book at its first level of occurrence to be ranked less reliably based on coursebooks. This could be because we are investigating MWEs rather than single items, which previous work has focused on. Future research, should continue to investigate different ways of linking items to levels and make sure to investigate both single words and MWEs. If empirical data is used as a basis, rather than e.g. crowdsourcing, both receptive and productive data should be investigated.

Compositionality (or transparency) proved to be problematic since the annotators did not appear to have interpreted the task in the same way. Still, by using the compositionality score from *one* of the annotators, we explore if there might be a tendency that relative rankings are linked to compositionality or transparency. Our results indicate that this might be the case, at least with regards to whether interjections or fixed expressions (group 1 in the crowdsourcing experiment) are interpreted as easier or more difficult than the first level of occurrence in coursebooks. This should be investigated further but it requires better ways of making sure whether it is transparency or compositionality that is estimated.

Some of the crowd rankings were also compared to occurrences in newspapers and blogs where the most difficult idiom was found to be rare in all of the corpora. Some further comparisons of this kind were done in a previous qualitative analysis (Lindström Tiedemann et al. 2022), see also Section 5.2.1. This showed that some MWEs may be more common in informal genres such as blogs, as claimed in Prentice & Sköldberg (2013).

We see tendencies for MWE usage to depend on many different factors including the tasks and genres which we compare. This means that we need resources such as the Swedish L2 Profiles which facilitates the comparison of texts aimed at learners and texts written by learners, and which also provides a possibility of easily comparing data to other corpora. We also have to remember that learner data is often quite small and often consists of fairly short texts (cf. Forsberg & Bartning 2010) which can complicate analysis and in particular comparisons with L1 production which often consists of longer texts and much larger corpora.

Making use of the data from the Swedish L2 Profiles categorised into MWE types, we see a clear development for MWEs in both receptive and productive data. The percentage of MWEs among the lemmagrams+sense types might possibly be an indicator of the productive proficiency level since there is a very clear increase per level and it is only at C1 that this is the same as in coursebooks at the group level, but also at B2 it is very close. Some subcategories show a clearer development (e.g. reflexive verbs, adverbial MWEs) whereas others seem fairly stable (e.g. support verbs, other contiguous MWEs such as idioms), and occasionally there is an indication of possible overuse. Future studies should try to investigate this further while restricting the genre and topic to match as much as possible, but also by investigating MWEs per text in both receptive and productive data in relation to the CEFR levels. However, this is complicated by the fact that learner texts tend to be rather short. This means that it is unlikely that several MWEs will be used in one text.

Acknowledgments

We would like to thank Riksbankens Jubileumsfond who financed most of the work presented in this paper through grant P17-0716:1. Some parts were also financed by the University of Helsinki and by Språkbanken Text, the University of Gothenburg.

We also wish to thank the editors and the anonymous reviewers for their valuable feedback on earlier versions of this chapter.

Abbreviations

CEFR	Common European Framework of Reference for Languages
GSE	Global Scale of English
L1	first language
L2	second language
MI	mutual information
MWE	multiword expression
NLP	natural language processing
NP	noun phrase
POS	part-of-speech
SAG	The Swedish Academy Grammar (Teleman et al. 1999)
SAOL	The Swedish Academy Glossary (Svenska Akademien 2015)

References

- Abrahamsson, Niclas & Kenneth Hyltenstam. 2009. Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning* 59(2). 249–306.
- af Hällström, Charlotta & Mikael Reuter. 2008. *Finlandssvensk ordbok*. Electronic version of the 4th edition. Schildts förlag & Forskningscentralen för de inhemska språken. <https://kaino.kotus.fi/fsob/>.
- Agirre, Eneko, Izaskun Aldezabal & Eli Pociello. 2006. Lexicalization and multiword expressions in the Basque WordNet. In Petr Sojka, Key-Sun Choi, Christine Fellbaum & Piek Vossen (eds.), *Proceedings of third international WordNet conference*, 131–138.
- Alfter, David. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. University of Gothenburg. (Doctoral dissertation).
- Alfter, David, Yuri Bizzoni, Anders Agebjörn, Elena Volodina & Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In Elena Volodina, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam & Lars Borin (eds.), *Proceedings of the joint Workshop on NLP for Computer-Assisted Language Learning and NLP for Language Acquisition (NLP4CALL & NLP4LA)*, 1–7. Umeå: Linköping University Electronic Press. <https://aclanthology.org/W16-6501>.
- Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2019. LEGATO: A flexible lexicographic annotation tool. In Mareike Hartmann & Barbara Plank (eds.), *22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 382–388. Linköping University Electronic Press.
- Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. *Northern European Journal of Language Technology* 7.
- Anward, Jan & Per Linell. 1976. Om lexikaliserade fraser i svenskan. *Nysvenska Studier* 55–56. 77–119.
- Attia, Mohammed, Antonio Toral, Lamia Tounsi, Pavel Pecina & Josef Van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In Éric Laporte, Preslav Nakov, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the 2010 workshop on multiword expressions: From theory to applications*, 19–27. Coling 2010 Organizing Committee.
- Baldwin, Timothy & Francis Bond. 2002. Multiword expressions: Some problems for Japanese NLP. In *Proceedings of the 8th annual meeting of the Association for Natural Language Processing*, 379–382. Keihanna, Japan.

- Bhalla, Vishal & Klara Klimcikova. 2019. Evaluation of automatic collocation extraction methods for language learning. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán & Torsten Zesch (eds.), *Proceedings of the 14th workshop on innovative use of NLP for building educational applications*, 264–274. Association of Computational Linguistics.
- Björklund, Siv. 2007. Reflexiva verb och infinitivfraser. *Språkbruk*. <https://www.sprakbruk.fi/-/reflexiva-verb-och-infinitivfraser>.
- Boers, Frank, June Eyckmans, Jenny Kappel, Hélène Stengers & Murielle Demecheleer. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research* 10(3). 245–261.
- Borin, Lars. 2021. Multiword expressions: A tough typological nut for Swedish FrameNet++. In Dana Dannélls, Lars Borin & Karin Friberg Heppin (eds.), *The Swedish FrameNet++: harmonization, integration, method development and practical language technology applications*, 221–260. Amsterdam/Philadelphia: John Benjamins.
- Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The sixth Swedish language technology conference (SLTC)*. Umeå University.
- Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47. 1191–1211.
- Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerch, Mehmet Doğan Uğur, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, 474–478. European Language Resources Association.
- Caines, Andrew & Paula Buttery. 2017. The effect of task and topic on opportunity of use in learner corpora. In Vaclav Brezina & Lynne Flowerdew (eds.), *Learner corpus research: New perspectives and applications*, 5–27. Bloomsbury Publishing Academic London.
- Capel, Annette. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal* 3.
- Capel, Annette. 2015. The English vocabulary profile. *English profile in practice* 5. 9–27.
- Cieślicka, Anna B. 2015. Idiom acquisition and processing by second/foreign language learners. In Roberto R. Heredia & Anna B. Cieślicka (eds.), *Bilingual figurative language processing*, 288–244. Cambridge University Press.

- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Modern Languages Division (Strasbourg) & Cambridge University Press.
- Council of Europe. 2009. *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Neus Figueras, Brian North, Sauli Takala, Piet Van Avermaet & Norman Verhelst (eds.). Council of Europe, Language policy division.
- Council of Europe. 2020. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Council of Europe Publishing.
- Cowie, Anthony P. 1994. Phraseology. In R. E. Asher (ed.), *Encyclopedia of language and linguistics*, 3168–3171. Oxford: Pergamon.
- De Cock, Sylvie, Sylviane Granger, Geoffrey Leech & Tony McEnery. 2014. An automated approach to the phrasicon of EFL learners. In Sylviane Granger (ed.), *Learner English on computer*, 67–79. Routledge.
- de Caseli, Helena Medeiros, Carlos Ramisch, Maria das Graças Volpe Nunes & Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation* 44. 59–77.
- Dürlich, Luise & Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Héléne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the eleventh international conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1140>.
- Durrant, Phil. 2018. Formulaic language in English for academic purposes. In Anna Siyanova-Chanturia & Ana Pellicer-Sánchez (eds.), *Understanding formulaic language: A second language acquisition perspective*, 211–227. Routledge.
- Ekberg, Lena. 1999. Användningen av komplexa predikat hos invandrarbarn i Rosengård. In Lars-Gunnar Andersson, Aina Lundqvist, Kerstin Norén & Lena Rogström (eds.), *Svenskans beskrivning 23: förhandlingar vid tjugotredje Sammankomsten för svenskans beskrivning: Göteborg den 15–16 maj 1998*, 86–95. Lund: Lund University Press.
- Ekberg, Lena. 2013. Grammatik och lexikon hos svenska i andraspråk på nästan infödd nivå. In Kenneth Hyltenstam & Inger Lindberg (eds.), *Svenska som andraspråk*, 259–279. Lund: Studentlitteratur.
- Ellis, Nick C. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics* 32. 17–44.

- Enström, Ingegerd. 1990. *Feltyper i invandrargymnasisters användning av partikelverb, prefixverb och reflexiva verb* (Rapporter från institutionen för nordiska språk/svenska (NORDRAPP) 4). Göteborg: Göteborgs universitet.
- Erman, Britt. 2007. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics* 12(1). 25–53.
- Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text & Talk* 20(1). 29–62.
- Forsberg, Fanny. 2006. *Le langage préfabriqué en français parlé L2: Étude acquisitionnelle et comparative*. Stockholm: Stockholm University. (Doctoral dissertation).
- Forsberg, Fanny. 2008. *Le langage préfabriqué: Formes, fonctions et fréquences en français parlé L2 et L1* (Contemporary Studies in Descriptive Linguistics 20). Oxford: Peter Lang.
- Forsberg, Fanny. 2010. Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching* 48. 25–51.
- Forsberg, Fanny & Inge Bartning. 2010. Can linguistic features discriminate between the communicative CEFR-levels? A pilot study of written L2 French. In Inge Bartning, Maisa Martin & Ineke Vedder (eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (EuroSLA Monographs 1), 133–158. EuroSLA.
- François, Thomas, Núria Gala, Patrick Watrin & Cédric Fairon. 2014. FLELex: A graded lexical resource for French foreign learners. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 3766–3773. Reykjavik, Iceland: European Language Resources Association (ELRA). <https://aclanthology.org/L14-1>.
- François, Thomas, Elena Volodina, Ildikó Pilán & Anaïs Tack. 2016. SVALex: A CEFR-graded lexical resource for Swedish foreign and second language learners. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC)*, 213–219. <https://aclanthology.org/L16-1>.
- Gala, Núria, Thomas François, Delphine Bernhard & Cédric Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In Phillipe Blache, Frédéric Béchet & Brigitte Bigi (eds.), *Proceedings of TALN 2014*, 91–102. Association pour le Traitement Automatique des Langues. <https://aclanthology.org/F14-2000>.

- Gala, Núria, Thomas François & Cédric Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets & Maria Tuulik (eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. proceedings of the eLex 2013 conference*, 132–151. <http://eki.ee/elex2013/proceedings/eLex2013-proceedings.pdf>.
- Granger, Syviane. 2018. Formulaic sequences in learner corpora. In Anna Siyanova-Chanturia & Ana Pellicer-Sanchez (eds.), *Understanding formulaic language: A second language acquisition perspective*, 228–247. Routledge.
- Green, Anthony. 2012. *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range* (English Profile Studies 2). Cambridge, UK: Cambridge University Press.
- Hawkins, John A & Luna Filipović. 2012. *Criterion features in L2 English: Specifying the reference levels of the Common European Framework*, vol. 1. Cambridge University Press.
- Jackendoff, Ray. 1997. *The architecture of the language faculty* (Linguistic Inquiry Monographs 28). Cambridge, MA: MIT Press.
- Karlsson, Ola. 2017. *Svenska skrivregler*. Liber.
- Krippendorff, Klaus. 1980. *Content Analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.
- Kurtes, Svetlana & Nick Saville. 2008. The English profile programme: An overview. *Research Notes* 33. 2–4.
- Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.
- Lewis, Margareta. 2008. *The idiom principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency, and proficiency*. Stockholm University. (Doctoral dissertation).
- Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2022. CEFR-nivåer och svenska flerordsuttryck. In Siv Björklund, Bodil Haagensen, Marianne Nordman & Anders Westerlund (eds.), *Svenskan i Finland 19: Föredrag vid den nittonde sammankomsten för beskrivningen av svenskan i Finland, Vasa den 6–7 maj 2021*, 218–233. Svensk-österbottniska samfundet.

- Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In *Constructicography: Constructicon development across languages*, 41–106. Amsterdam/Philadelphia: John Benjamins. DOI: 10.1075/cal.22.03lyn.
- Mel'čuk, Igor. 1998. Collocations and lexical functions. In A. P. Cowie (ed.), *Phraseology: Theory, analysis, and applications*, 23–53. Clarendon Press.
- Meunier, Fanny. 2012. Formulaic language and language teaching. *Annual Review of Applied Linguistics* 32. 111–129.
- Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2). 223–242.
- Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.
- O'Keeffe, Anne & Geraldine Mark. 2017. The English grammar profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics* 22(4). 457–489.
- Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research* 35(1). 121–145.
- Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & R. W. Schmidt (eds.), *Language and communication*, 203–239. Routledge.
- Piao, Scott Songlin, Paul Rayson, Dawn Archer & Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language* 19(4). 378–397.
- Prentice, Julia. 2010. *Käppen i hjulen: Behärskning av svenska konventionaliserade uttryck bland gymnasieelever med varierande språklig bakgrund* (ROSA 12). University of Gothenburg. <http://hdl.handle.net/2077/23261>.
- Prentice, Julia & Emma Sköldböck. 2013. Flerordsenheter – ur ett andraspråksperspektiv. In Kenneth Hyltenstam & Inger Lindberg (eds.), *Svenska som andraspråk: I forskning, undervisning och samhälle*, 2nd edn., 197–220. Studentlitteratur.
- Ringbom, Håkan. 2012. A country in focus: Review of recent applied linguistics research in Finland and Sweden, with specific reference to foreign language learning and teaching. *Language Teacher* 45(4). 490–514.
- Rudebeck, Lisa, Gunlög Sundberg & Mats Wirén. 2021. *SweLL normalization guidelines* (GU-ISS Research report series). University of Gothenburg. <http://hdl.handle.net/2077/69432>.

- Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander F. Gelbukh (ed.), *Proceedings of the third international conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, 1–15. Springer.
- Schmidt, Richard. 2012. Attention, awareness, and individual differences in language learning. In Wai Meng Chan, Kwee Nyet Chin, Sunil Kumar Bhatt & Izumi Walker (eds.), *Perspectives on individual characteristics and foreign language education* (SSFLE 6), 27–50. De Gruyter Mouton. DOI: 10.1515/9781614510932.27.
- Schulte im Walde, Sabine. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 269–308. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998645.
- Shigeto, Yutaro, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung & Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In Valia Kordoni, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the 9th workshop on multiword expressions (MWE 2013)*, 139–144. Atlanta, GA: Association for Computational Linguistics.
- Sköldberg, Emma, Linnéa Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Leif-Jöran Olsson, Julia Prentice, Rudolf Rydstedt, Sofia Tingsell & Jonatan Uppström. 2013. Between grammars and dictionaries: A Swedish constructicon. In Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets & Maria Tuulik (eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 conference*, 310–327. http://eki.ee/elex2013/proceedings/eLex2013_21_Skoldberg+etal.pdf.
- Svenska Akademien. 2015. *Svenska Akademiens ordlista*. 14th edn. Svenska Akademien & Norstedts ordbok. <https://www.svenskaakademien.se/svenska-spraket/svenska-akademiens-ordlista-saol>.
- Svenska Akademien. 2021. *Svensk ordbok utgiven av Svenska Akademien*. 2nd edn. Svenska Akademien & Nordstedts ordbok. <https://www.svenskaakademien.se/svenska-spraket/svensk-ordbok-utgiven-av-svenska-akademien-so>.
- Tack, Anaïs, Thomas François, Piet Desmet & Cédric Fairon. 2018. NT2Lex: A CEFR-graded lexical resource for Dutch as a foreign language linked to open Dutch WordNet. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock & Helen Yannakoudakis (eds.), *Proceedings of the 13th workshop on innovative use of NLP for Building Educational Applications (BEA)*, 137–146. New

- Orleans, LA: Association for Computational Linguistics. <https://aclanthology.org/W18-05>.
- Teleman, Ulf, Staffan Hellberg, Erik Andersson & Lisa Christensen. 1999. *Svenska akademiens grammatik*. Norstedts ordbok & Svenska Akademien.
- Villada Moirón, María Begoña. 2005. *Data-driven identification of fixed expressions and their modifiability*. University of Groningen. (Doctoral dissertation).
- Volodina, Elena, David Alfter & Therese Lindström Tiedemann. 2022a. Crowdsourcing ratings for single lexical items: A core vocabulary perspective. *Slovenščina 2.0* 10(2). 5–61. DOI: 10.4312/slo2.0.2022.2.5-61.
- Volodina, Elena, David Alfter, Therese Lindström Tiedemann, Maisa Lauriala & Daniela Piipponen. 2022b. Reliability of automatic linguistic annotation: Native vs non-native texts. In Monica Monachini & Maria Eskevich (eds.), *Selected papers from the CLARIN Annual Conference 2021*, 151–167. Linköping Electronic Press. DOI: 10.3384/9789179294441.
- Volodina, Elena, Ildikó Pilán, Stian Rødven Eide & Hannes Heidarsson. 2014. You get what you annotate: A pedagogically annotated corpus of coursebooks for Swedish as a second language. In Elena Volodina, Lars Borin & Ildikó Pilán (eds.), *Proceedings of the 3rd workshop on NLP for computer-assisted language learning (NLP4CALL)*, 128–144. <https://aclanthology.org/W14-3500>.
- Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg & Monica Sandell. 2016a. Swell on the rise: Swedish learner language corpus for European reference level studies. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Héléne Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC'16)*, 206–212. Portorož, Slovenia: ACL. <https://aclanthology.org/L16-1>.
- Volodina, Elena, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse & Thomas François. 2016b. SweLLex: Second language learners' productive vocabulary. In Elena Volodina, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam & Lars Borin (eds.), *Proceedings of the joint Workshop on NLP for Computer-Assisted Language Learning and NLP for Language Acquisition (NLP4CALL & NLP4LA)*, 76–84. Umeå: Linköping Electronic Conference Proceedings. <https://aclanthology.org/W16-6500>.
- Watrín, Patrick & Thomas François. 2011. An n-gram frequency database reference to handle MWE extraction in NLP applications. In Valia Kordoni, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the workshop on multiword expressions: From parsing and generation to the real world*, 83–91. Portland, OR:

- Association for Computational Linguistics. <https://aclanthology.org/W11-0813>.
- Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge University Press.
- Wray, Alison. 2013. Formulaic language. *Language teaching* 46(3). 316–334.
- Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20(1). 1–28.
- Yamaguchi, Nami, David Alfter, Kaori Sugiyama & Thomas François. 2022. Towards a verb profile: Distribution of verbal tenses in FFL textbooks and in learner productions. In David Alfter, Elena Volodina, Thomas François, Piet Desmet, Frederik Cornillie, Arne Jönsson & Evelina Rennes (eds.), *Proceedings of the 11th workshop on NLP for computer assisted language learning*, 123–142. Louvain-la-Neuve, Belgium: LiU Electronic Press. <https://aclanthology.org/2022.nlp4call-1.13>.

Name index

- Abrahamsson, Niclas, 310, 313
Adamou, Evangelia, 42, 44
af Hällström, Charlotta, 335
Agirre, Eneko, 312
Aijmer, Karin, 9
Al-Haj, Hassan, 55, 77, 79, 81
Alfter, David, 311, 315, 323–326, 329,
337
Alipoor, Pegah, 271, 278, 279, 295
Anastasiadis-Symeonidis, Anna, 153,
162
Anthony, Laurence, 159
Anward, Jan, 316
Attia, Mohammed, 312
Augustinus, Liesbeth, 230, 237, 260
Autelli, Erica, 76, 80, 81

Baccianella, Stefano, 100
Baker, Collin F., vii, 150
Baldwin, Timothy, 3, 4, 17, 31, 49, 147,
270, 312
Barbu Mititelu, Verginica, 82, 84, 89,
102, 103, 109
Baroni, Marco, 272
Bartning, Inge, 313, 314, 318, 338
Bauer, Laurie, 270
Bechhofer, Sean, 191
Bell, Melanie J., 273
Benczes, Réka, 270
Bettinger, Julia, 270
Bhalla, Vishal, 311, 312
Bieliková, Mária, 215

Bildhauer, Felix, 272, 274, 278
Björklund, Siv, 335
Boas, Hans C., 150
Boers, Frank, 314
Bond, Francis, 312
Bonial, Claire, 155
Borgwaldt, Susanne, 270, 273
Borin, Lars, 150, 156, 316, 318, 319,
323, 324, 327, 333
Bosque-Gil, Julia, 193
Bott, Stefan, 278
Bouma, Gosse, 237
Bozdéčová, Ivana, 3
Brač, Ivana, 152
Brenzinger, Matthias, 42
Bresnan, Joan, 75
Brewer, Ebenezer Cobham, 199
Brin, Sergey, 222
Broekhuis, Hans, 249
Brouwer, Matthijs, 261
Brugman, Hennie, 232
Brysbart, Marc, 288
Buchholz, Sabine, 124
Buendía Castro, Miriam, 152
Burchardt, Aljoscha, 156
Burger, Harald, 3
Butterworth, Brian, 270
Buttery, Paula, 334

Caines, Andrew, 334
Calzolari, Nicoletta, vi, 153
Candito, Marie, 150, 157, 159

Name index

- Cap, Fabienne, 270
Capel, Annette, 314
Carpuat, Marine, 270
Castellan, N. John, 278
Čechová, Marie, 4
Čermák, František, 2–4, 7, 8
Chiarcos, Christian, 75, 189, 190, 195,
196, 208, 216
Cholakov, Kostadin, 270
Chu, Yen-Lun, 215
Cieślicka, Anna B., 315, 324
Cimiano, Philipp, 192
Clouet, Elizaveta Loginova, 270
Colson, Jean-Pierre, 4
Constant, Mathieu, v, 230
Constantinides, Nicolaos Th., 42, 44
Cook, Paul, 278
Copestake, Ann, vi, 153
Cordeiro, Silvio, 271–273, 278
Costello, Fintan J., 270
Coulmas, Florian, 9
Council of Europe, 310, 313, 315, 338
Cowie, Anthony P., 319

Daille, Béatrice, 270
Dalrymple, Mary, 75
Dankers, Verna, 270
de Caseli, Helena Medeiros, 312
De Cock, Sylvie, 313
de Does, Jesse, 232, 261
de Jong, Nicole H., 275
de Marneffe, Marie-Catherine, 58, 89
Declerck, Thierry, 208
Di Fabio, Andrea, 154
Diab, Mona, 270
Dima, Corina, 278
Dolbey, Andrew, 152
Dürlich, Luise, 315
Durrant, Phil, 312

Dyvik, Helge, 75–77, 79, 80, 119

Eichel, Annerose, 270
Ekberg, Lena, 313, 318
Elhadad, Michael, 150
Ellis, Nick C., 294, 313
Ellsworth, Michael, 150, 156
Enström, Ingegerd, 310, 318, 335, 337
Erk, Katrin, 150
Erman, Britt, 313, 318
Evert, Stefan, v, 189

Faber, Pamela, 152
Fanciullo, Davide, 42, 44
Feldweg, Helmut, 208, 275, 279
Fellbaum, Christiane, vi, vii, 75–77,
79–83, 87, 100, 101, 154, 273,
279
Filipović, Luna, 314
Fillmore, Charles J., vii, 149
Finkbeiner, Rita, 189
Firth, John R., 277
Forsberg, Fanny, 310, 313, 314, 318,
338
Forster, Kenneth I., 270
Fotopoulou, Aggeliki, 81, 153, 154
François, Thomas, 312, 315, 323
Francopoulo, Gil, 192
Fried, Mirjam, 75

Gagné, Christina L., 270, 281
Gala, Núria, 315
Gamallo, Pablo, 270
Gantar, Polona, 51
Gavriilidou, Zoe, 162
Geyken, Alexander, 75–77, 79–81, 87,
100
Giouli, Voula, 57, 75, 80, 148, 150, 152,
155, 157, 165, 173

- Girju, Roxana, 270
Goldhahn, Dirk, 208
Gracia, Jorge, 193
Granger, Sylviane, 312
Green, Anthony, 314
Grégoire, Nicole, 51, 54, 56, 75, 76, 78,
79, 81, 118, 154, 233
Gross, Gaston, 77
Gross, Maurice, 75, 79, 147, 153
Groß, Thomas, 124
Grün, Christian, 259

Hajič, Jan, 31
Hamp, Birgit, 208, 275, 279
Hardie, Andrew, 189
Harris, Zellig, 277
Hartmann, Silvana, 154
Haspelmath, Martin, 58
Hätty, Anna, 270
Hawkins, John A, 314
Hayoun, Avi, 150
Hermann, Karl Moritz, 278
Hnátková, Milena, 2, 4, 11, 29, 32, 76–
79
Hoekstra, Heleen, 248
Holton, David, 165, 168
Hüllen, Werner, 57
Hüning, Matthias, 188
Hyltenstam, Kenneth, 310, 313

Iñurrieta, Uxoá, 75

Jackendoff, Ray, 312, 319, 333
Jelínek, Tomáš, 4, 12, 25, 28, 31
Jespersen, Otto, 49
Joshi, Pratik, 56

Karahóga, Ritván, 42, 43, 56
Karahóga, Sebjajdín, 42
Karlsson, Ola, 317

Keane, Mark T., 270
Kilgarriff, Adam, 159, 189, 208
Kim, Jeong-uk, 150
Kim, Su Nam, 3, 4, 17, 31, 49, 147, 270
Kipper, Karin, 154
Klégr, Aleš, 4
Klimcikova, Klara, 311, 312
Klimek, Bettina, 196
Kochová, Pavla, 32
Koehn, Philipp, 242
Koeva, Svetla, 74, 82, 84, 109
Kokkas, Nikolaos, 42
Kompan, Michal, 215
Köper, Maximilian, 271, 278, 279, 295
Kopřivová, Marie, 2
Kordoni, Valia, 270
Kovářík, Oleg, 3
Kováříková, Dominika, 3, 4
Krimpas, Panagiotis G., 44, 50
Krippendorff, Klaus, 328
Kuiper, Koenraad, 154
Kunze, Claudia, 275, 279
Kurtes, Svetlana, 314

Lapata, Mirella, 278
Laporte, Éric, 49, 153
Larose, Chantal D., 221, 222
Larose, Daniel T., 221, 222
Laskova, Laska, 117, 121
Lenci, Alessandro, 150
Lendvai, Piroška, 208
Leseva, Svetlozara, 3, 56, 57, 74, 102,
103, 120, 154, 321
Levi, Judith N., 270, 275
Lewis, Margareta, 318
Libben, Gary, 273
Lichte, Timm, vi, 4, 54, 78, 107, 108,
118
Lindén, Krister, 150

Name index

- Lindström Tiedemann, Therese, 311,
326, 330, 338
Linell, Per, 316
Lion-Bouton, Adam, 119
Liu, Kaiying, 150
Lopatková, Markéta, 3
Losnegaard, Gyri Smørdal, 51
Lyngfelt, Benjamin, 328
Lyons, John, 57
- Ma, Xuezheng, 10, 31
Machálek, Tomáš, 28
Manning, Christopher D., 220, 221
Mark, Geraldine, 314
Markantonatou, Stella, 40, 42, 51, 53,
59, 75–81, 154
Marsh, Charles, 280
Marsi, Erwin, 124
Martins, André, 10
Masini, Francesca, 119
McCrae, John Philip, 187, 190, 192
Mel'čuk, Igor, 75, 77–80, 188, 313, 319
Meunier, Fanny, 314
Mikolov, Tomas, 30
Miletic, Filip, 271, 278, 279, 288, 295
Miller, George A., vi, 82, 101
Mini, Marianna, 153
Mitchell, Jeff, 278
Molich, Rolf, 59
Monti, Johanna, 75, 78, 80
Moon, Rosamund, 4
Müller, Stefan, 75
Muraki, Emiko J., 288
Murphy, Gregory L., 270
- Nadif, Mohamed, 219, 220
Nastase, Viviana A., 270
Navigli, Roberto, 108
Nesselhauf, Nadja, 313, 314, 319
- Ní Loingsigh, Katie, 40
Nielsen, Jakob, 59
Nissim, Malvina, 153
Nivre, Joakim, 154, 260
Nunberg, Geoffrey, 148, 315, 324
- Ó Raghallaigh, Brian, 40
Ó Séaghdha, Diarmuid, 275, 278
O'Grady, William, 121, 124
O'Keefe, Anne, 314
Odijk, Jan, 75–77, 79–81, 153, 231, 233,
237, 248
Ohara, Kyoko, 150
Oostdijk, Nelleke, 258
Opavská, Zdenka, 32
Ordelman, Roeland J.F., 257, 258
Osborne, Timothy, 124, 130
Osenova, Petya, 10, 56, 78–81, 117,
121, 122, 136, 154
Osherson, Anne, vii
Östman, Jan-Ola, 75
Ostroški Anić, Ana, 152
- Palmer, Martha, 155
Papadimitriou, Panayotis, 42
Paquot, Magali, 309, 312
Pasquer, Caroline, 3
Pawley, Andrew, 310, 313
Pedersen, Ted, 220
Pergl, Robert, 196
Perkins, Michael R., 312
Petkevič, Vladimír, 29
Petrucek, Miriam R. L., 149, 156
Piao, Scott Songlin, 312
Piirainen, Elisabeth, 40, 50
Pilitsidou, Vera, 148, 150, 152
Plag, Ingo, 270
Pollard, Carl, 75
Ponzetto, Simone Paolo, 108

- Popovičová, Snežana, 3
Prentice, Julia, 310, 313, 316, 317, 330, 338
Przepiórkowski, Adam, 3, 75, 78, 79, 118

Ralli, Angela, 162
Ramisch, Carlos, v, x, 155, 165
Reddy, Siva, 271–273, 278
Reuter, Mikael, 335
Ringbom, Håkan, 313
Role, François, 219, 220
Roller, Stephen, 270, 278
Romary, Laurent, 211
Rosen, Alexandr, 10
Rosetta, M. T., 252
Rudebeck, Lisa, 325
Ruppenhofer, Josef, 150, 157
Rychlý, Pavel, 220

Sag, Ivan A., 75, 148, 188, 231, 270, 312, 316, 319, 321
Sager, Juan C., 152
Sailer, Manfred, 53
Saito, Hiroaki, 150
Salehi, Bahar, 270, 278
Salomão, Maria Margarida M., 150
Sandry, Susan, 42
Savary, Agata, v, 47, 49, 55, 56, 58, 73, 87, 89, 120, 154, 155, 165, 166, 170, 270
Saville, Nick, 314
Schäfer, Martin, 273
Schäfer, Roland, 272, 274, 278
Schafroth, Elmar, 75
Schlücker, Barbara, 188, 189
Schmidt, Richard, 314
Schmidt, Thomas C., 152
Schneider, Nathan, 120, 155

Schulte im Walde, Sabine, 20, 270–275, 278, 279, 288, 294, 295, 315
Schütze, Hinrich, 220, 221
Sheinfux, Livnat Herzig, 4
Shigeto, Yutaro, 312
Shudo, Kosho, 75–77, 79, 81, 153
Siegel, Sidney, 278
Simov, Kiril, 10, 56, 78–81, 117, 121, 136, 154
Sköldberg, Emma, 310, 316, 317, 328, 330, 338
Skoumalová, Hana, 10, 54, 56, 57, 75, 76, 78–81, 119
Smolka, Eva, 270
Spalding, Thomas L., 270
Stoett, Frederik August, 233
Straka, Milan, 89, 167
Straková, Jana, 167
Subirats, Carlos, 150
Suchánek, Marek, 196
Svenska Akademien, 317, 332, 339
Syder, Frances Hodgetts, 310, 313

Tack, Anaïs, 315
Taft, Marcus, 270
Tasovac, Toma, 212
Tayyar Madabushi, Harish, 155
Teleman, Ulf, 316, 339
Temmerman, Rita, 3
Text Encoding Initiative, 212
Theocharides, Petros, 42
Timponi Torrent, Tiago, 150, 152
Tufiş, Dan, 82, 84

Urešová, Zdeňka, 3

van de Camp, Matje, 232
van der Beek, Leonoor, 237

Name index

- Van Eynde, Frank, 247, 248
van Noord, Gertjan, 230, 233, 242,
248, 255, 258, 261
Venturi, Giulia, 152
Vietri, Simonetta, 75, 80
Villada Moirón, María Begoña, 312
Villavicencio, Aline, 54, 75–80, 153
Virk, Shafqat Mumtaz, 157
Volodina, Elena, 311, 315, 318, 319,
322–325, 327–329
von der Heide, Claudia, 273
Vondřička, Pavel, 22, 27, 52, 54, 56,
119
Voyatzi, Stavroula, 153

Wang, Tzone-I, 215
Warren, Beatrice, 318
Watrin, Patrick, 312
Weller, Marion, 270
Wisniewski, Edward J., 270, 280
Wray, Alison, 310, 312, 313

Yamaguchi, Nami, 315
Yimam, Seid Muhie, 167
Yong, Zheng Xin, 157
You, Liping, 150

Zampieri, Nicolas, 120
Zaninello, Andrea, 153

Multiword expressions in lexical resources

This volume contains chapters that paint the current landscape of the multiword expressions (MWE) representation in lexical resources, in view of their robust identification and computational processing. Both large-size general lexica and smaller MWE-centred ones are included, with special focus on the representation decisions and mechanisms that facilitate their usage in Natural Language Processing tasks. The presentations go beyond the morpho-syntactic description of MWEs, into their semantics.

One challenge in representing MWEs in lexical resources is ensuring that the variability along with extra features required by the different types of MWEs can be captured efficiently. In this respect, recommendations for representing MWEs in mono- and multi-lingual computational lexicons have been proposed; these focus mainly on the syntactic and semantic properties of support verbs and noun compounds and their proper encoding thereof.