Multiword expressions in lexical resources

Linguistic, lexicographic, and computational perspectives

Edited by

Voula Giouli

Verginica Barbu Mititelu

Phraseology and Multiword Expressions 6

language science press

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.

ISSN: 2625-3127

Multiword expressions in lexical resources

Linguistic, lexicographic, and computational perspectives

Edited by

Voula Giouli

Verginica Barbu Mititelu

language science press

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

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: XƎLATEX

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

Storage and cataloguing done by FU Berlin

Contents

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 wordnets[1] 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 Vondřička, 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

[1]The 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. Seul, 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 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]

[a]Applied Computational Linguistics, University of Augsburg, Germany [b]Institute for Digital Humanities, University of Cologne, Germany [c]Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest [d]Institute of Language and Speech Processing, Athena Research Center, Athens, Greece [e]Computational and Corpus Linguistics, University of Erlangen-Nuremberg, Germany [f]Consiglio 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.10998641

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*

[1]The 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 WorkBench[2] (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** (Chiarcos 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/

[8] Note 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

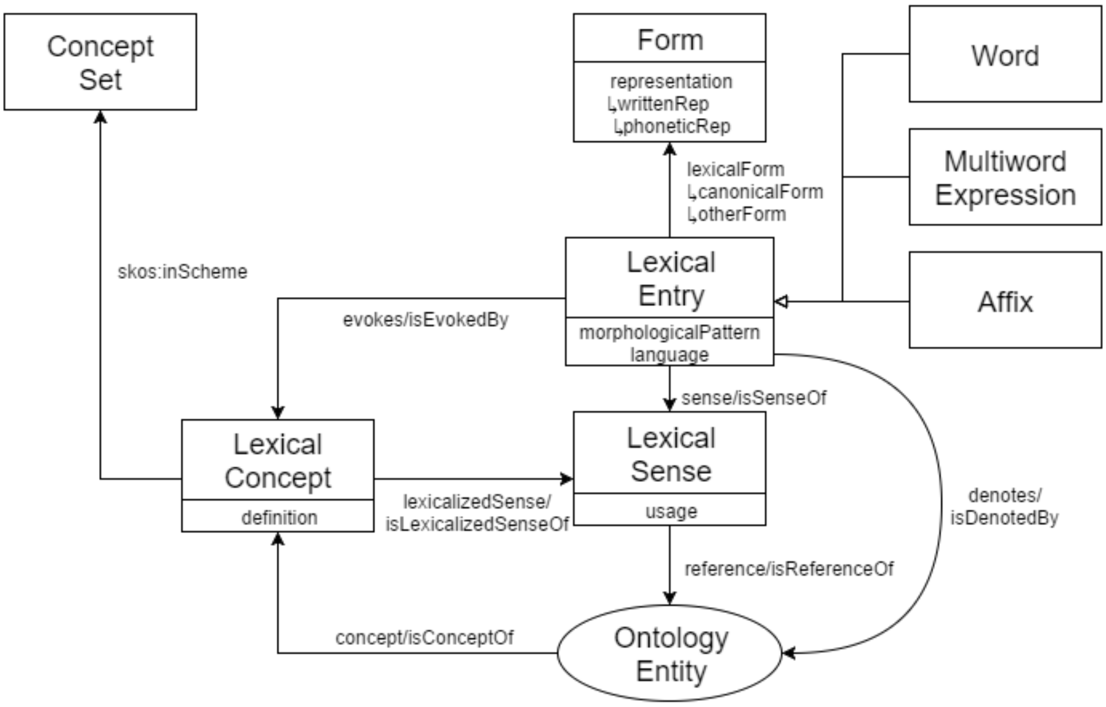

Figure 1: OntoLex-Core.

OntoLex-Core[9] (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

[10] In 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.

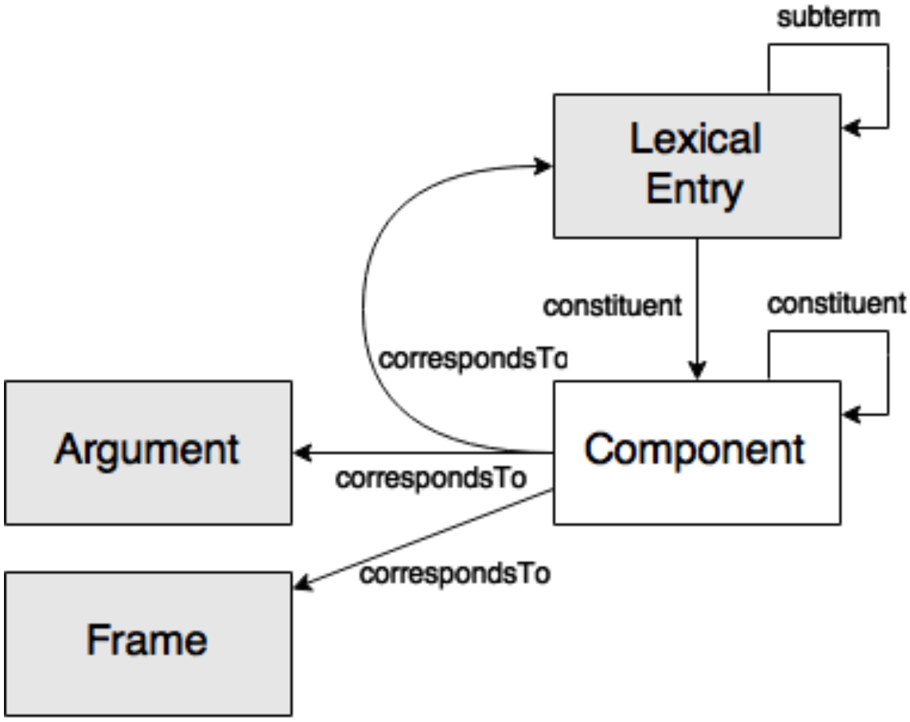

Figure 2: The OntoLex decomp module.

`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]

[12]This 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.

[13]For specific collocation metrics within **FrAC** see Appendix A.

[14]The 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**.

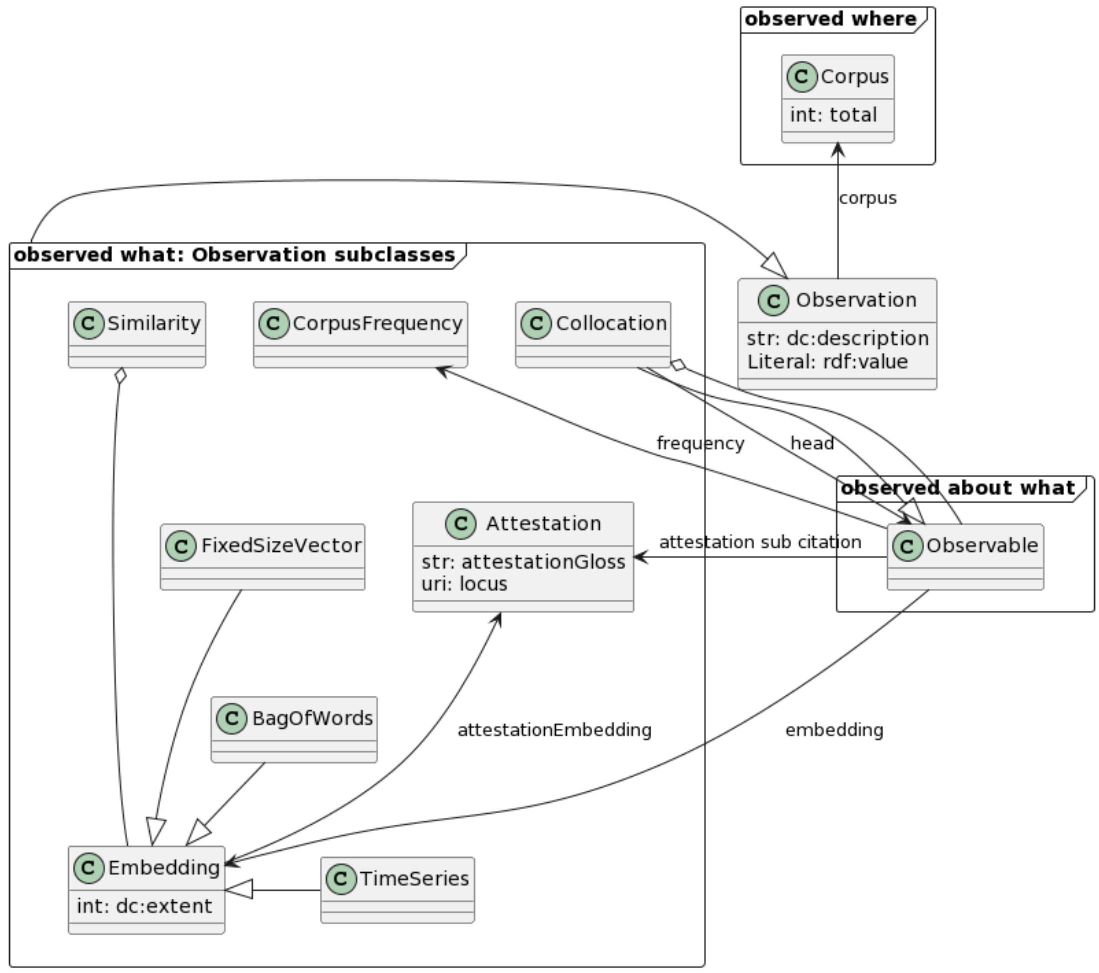

Figure 3: The OntoLex-FrAC module as an UML class diagram (see Suchánek & Pergl (2020) for notation), version July 2022.

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.

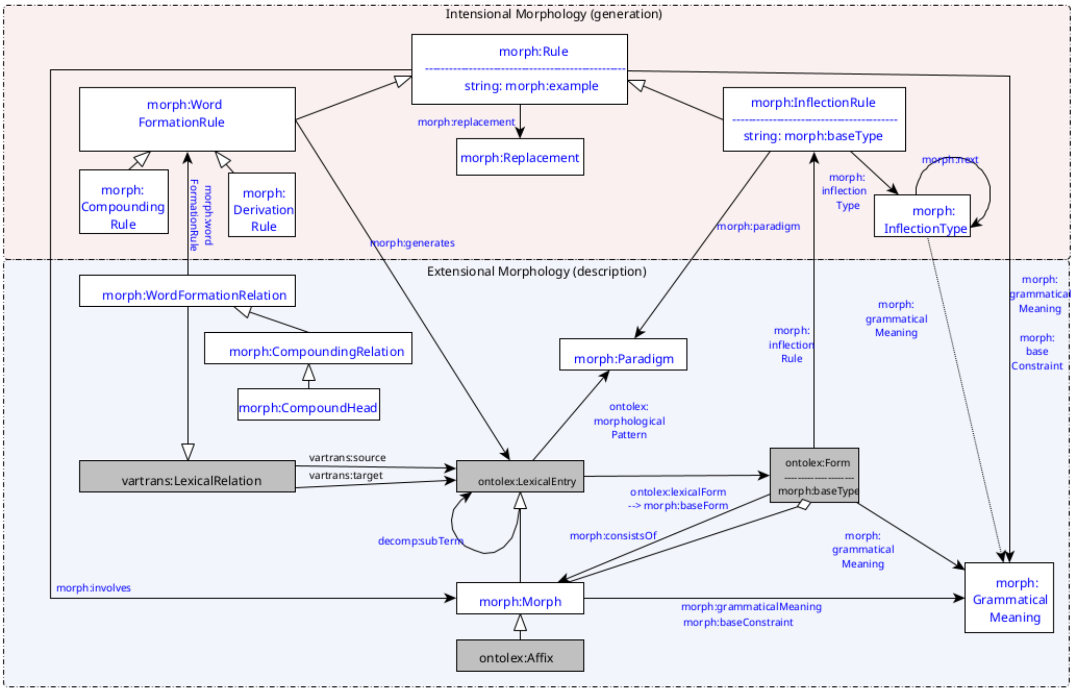

Figure 4: The OntoLex-Morph module, version 4.18 (October 2023).

OntoLex morphs are the central elements of the `morph:WordFormationRules` and `morph:InflectionRules` that involve them. Both types of rules can be defined by a `morph:example` (a string for descriptive morphology) or a `morph:replacement` (replacement pattern). The characteristic of word formation rules is that they describe a *lexical* process that creates an instance of `ontolex:LexicalEntry`. While a word formation rule formulates or illustrates a general pattern, the lexico-semantic relation between two specific lexical entries (such as the base and a derived word, or a constituent word and a compound) is modelled as `morph:WordFormationRelation`. In the case of compounding, the head can be made explicit using `morph:CompoundHead`. If no head is marked, one can use either `morph:CompoundingRelation` or **decomp**.

3 Modelling multiword expressions in OntoLex

As the reader will no doubt have appreciated by now, the **OntoLex-Core** vocabulary is not sufficient in and of itself for the task of describing how MWEs are formed and limits itself to allowing users to flag lexical entries as MWEs. We can make up for these expressive shortcomings, however, by availing ourselves 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 .
```

[15] We 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.

[16] Note 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/.

[17] We 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

```
:_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 SPARQL[20] 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 ] .
```

[19] Although 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

[21] Note 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 Books[22] 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 $p(\langle\texttt{:cat\_lex},\texttt{:eye\_lex}\rangle|\texttt{:eye\_lex})$. 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.

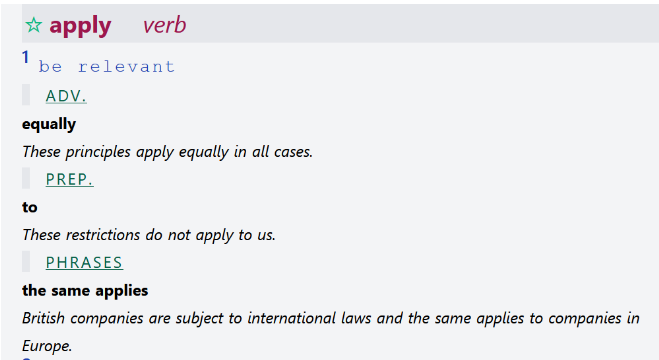

Figure 5: OZDIC: example *apply* (verb).

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 ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:verb ;
  ontolex:sense oz:apply-v-sense1 ;
  ontolex:canonicalForm [ ontolex:writtenRep "apply"@en ] .
oz:apply-v-sense1 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

[25]It 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.

criterion	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

[27] Although it would be better here to speak of *families* of models for lexical resources.

[28] LMF 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, renditional, 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

[29]Note 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

[33]In 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 *ab*using 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 GRDDL[37] or over JSON data with JSON-LD[38] 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/

[39]Queries 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 GlobaLex 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).

[40]This 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 simple 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:

$$\texttt{relFreq}_x = \frac{p(x,y)}{p(x)}$$

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:

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

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):

$$\mathrm{PMI}^2(x,y) = \log \frac{p(x,y)^2}{p(x)p(y)}$$

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):

$$\mathrm{PMI}^3(x,y) = \log \frac{p(x,y)^3}{p(x)p(y)}$$

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):

$$\mathrm{PMI}^k(x,y) = \log \frac{p(x,y)^k}{p(x)p(y)}$$

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):

$$\mathrm{NPMI}(x, y) = \frac{\mathrm{PMI}(x, y)}{-\log p(x, y)}$$

Definition 6.7 (`frac:pmiLogFreq`). The PMI log Freq (also know as Salience) is defined as:[41]

$$\text{PMI-logFreq}(x, y) = \mathrm{PMI}(x, y) \cdot \log(N p(x, y) + 1)$$

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):

$$\mathrm{Dice}(x, y) = \frac{2p(x, y)}{p(x) + p(y)}$$

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]

$$\mathrm{LogDice}(x, y) = 14 + \log_2 \mathrm{Dice}(x, y) = 14 + \log_2 \frac{2p(x, y)}{p(x) + p(y)}$$

Definition 6.10 (`frac:minSensitivity`). Minimum sensitivity is a measure of dependen-ce between word x and word y and it is computed as the minimum of the relative sensitivity of each word (Pedersen 1998):

$$\mathrm{minSensitivity}(x, y) = \min\left(\frac{p(x, y)}{p(y)}, \frac{p(x, y)}{p(x)}\right)$$

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:tscore`). The Student's t test (T-score) finds words whose co-occur-rence 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 that 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:cscore` 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):

$$\texttt{support}(x \rightarrow y) = p(x,y)$$

Definition 6.15 (`frac:confidence`). Confidence measures the probability of a rule to be true (Larose & Larose 2014):

$$\texttt{confidence}(x \rightarrow y) = \frac{p(x, y)}{p(x)}$$

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):

$$\texttt{lift}(x \rightarrow y) = \frac{p(x, y)}{p(x)p(y)}$$

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 {
```

```
    { 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 Truică. 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 Truică. 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. LexInfo: 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 Monachini, 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 Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ & 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-0. *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.

Name index

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 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 thereof.