

The Challenges of Syntactic Descriptions of Multiword Expressions in Electronic Lexicography

Verginica Barbu Mititelu¹, Voula Giouli², Gražina Korvel³, Chaya Liebeskind⁴, Irina Lobzhanidze⁵, Rusudan Makhachashvili⁶, Stella Markantonatou⁷, Alexandra Markovic⁸, Ivelina Stoyanova⁹

¹Romanian Academy Research Institute for Artificial Intelligence, Bucharest

²Aristotle University of Thessaloniki, Thessaloniki

³Vilnius University, Vilnius

⁴Jerusalem College of Technology Jerusalem, Israel

⁵Ilia State University, Tbilisi

⁶Borys Grinchenko Kyiv Metropolitan University, Kyiv

⁷Institute for Language and Speech Processing and Archimedes/Athena RC, Athens

⁸Institute for Serbian Language, Belgrade

⁹Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia

E-mail: vergi@racai.ro, pgiouli@del.auth.gr, grazina.korvel@mif.vu.lt, liebchaya@gmail.com, irina_lobzhanidze@iliauni.edu.ge, r.makhachashvili@kubg.edu.ua, marks@athenarc.gr, aleksandra.markovic@isj.sanu.ac.rs, iva@dcl.bas.bg

Abstract

In this paper, we provide a comprehensive overview of the way in which the morpho-syntactic properties of multiword expressions are represented in lexical resources to support Natural Language Processing downstream applications. Starting from an up-to-date and comprehensive overview of the existing lexica dedicated to multiword expressions and containing their syntactic description, we outline the current state of play in encoding syntactic information about multiword expressions (internal structure, argument structure, word order, discontinuity, verb alternations). We also discuss the relevance of the syntactic description of multiword expressions for several Natural Language Processing tasks. Our work contributes to the literature that fosters improvements in both the development and deployment of multiword expression lexica to ensure that they can support future Natural Language Processing innovations more effectively.

Keywords: multiword expression (MWE); lexica; morpho-syntactic description, Natural Language Processing

1. Introduction

Multiword expressions (MWEs) present significant challenges in Natural Language Processing (NLP) (Sag et al., 2002), primarily due to their semantic non-compositionality, though they have other idiosyncrasies that are also challenging (Baldwin & Kim, 2010), among them, the syntactic ones: irregular internal structure, combinatorial possibilities of the expression as a whole, constraints on the modification of components, constraints on syntactic transformations, word order variation, etc. Despite recent advances in NLP, including the advent of large language models (LLMs), as well as small ones, the inherent complexity and distributional properties of MWEs continue to impede their effective automatic processing. In this paper, we explore the current state-of-the-art in MWE lexicon

development, aiming to provide a comprehensive overview of how their morpho-syntactic properties are represented to serve the needs of NLP downstream applications. The focus on syntactic information is motivated by the fact that prior studies and lexicon developers alike have somewhat neglected syntax.

The paper is structured as follows. Section 2 outlines the main goals and scope of the survey. Section 3 briefly presents previous surveys, highlighting the differences between them and the current one. The methodology employed for collecting relevant resources and selecting the most appropriate ones for analysis is outlined in Section 4, where we also show a summary of the distribution of the selected resources with respect to the representation of syntactic information. Section 5 focuses in detail on how syntactic properties of MWEs are encoded across different computational lexica. The analysis includes PoS encoding, internal structure of the MWEs, argument structure, word order variations, modifications, discontinuity, and alternations. This is followed by Section 6, which discusses the relevance of the syntactic description of MWEs for various NLP tasks. Finally, Section 7 offers concluding remarks and outlines directions for future work.

2. Objectives and scope of the survey

Over the last few decades, a growing number of lexical resources have been developed to encode MWEs, capturing not only their surface forms but also their lexical, syntactic, semantic, and pragmatic properties. These resources aim to support both linguistic research and NLP applications by providing structured information about idioms, collocations, light verb constructions, and other fixed or semi-fixed expressions that may deviate from compositional norms.

The primary objective of this survey is to provide an overview of the ways in which syntactic properties of MWEs are represented in lexica. More specifically, it examines how these resources capture their internal structure, syntactic variability, argument realisation patterns, and how such information is integrated into syntactic models for NLP applications.

To the best of our knowledge, this is the first survey to systematically examine the particularities of MWE lexica in terms of encoding their syntactic properties. While previous work has addressed general aspects of MWE representation or focused on specific resource types or languages, our contribution lies in providing a cross-resource, syntax-oriented analysis that highlights how syntactic information is modeled, structured, and made available for computational use.

3. Related work

Several papers present the results of surveys on various aspects of the linguistic description of MWEs. Rosen et al. (2015) focused on how MWEs are syntactically treated in several treebanks and highlighted the need for harmonizing the annotation principles for facilitating the comparability of the phenomena between languages, which still holds even today, 10 years after their work.

Losnegaard et al. (2016) surveyed the language resources that contain MWEs (up to the year 2016). The syntactic description of MWEs is considered in their paper, as they include (10) treebanks with annotated MWEs among the resources they analyze, specifying the

syntactic framework adopted (dependency, dependency and constituents hybrid annotation, or functional), the fact that non-contiguous MWEs are annotated, and annotated MWE types (idioms, phrasal verbs, light verb constructions, named entities, compounds, etc.). Concerning MWE lexica, the authors discuss two syntax-related aspects: contiguity and valence frames of MWEs. Encoding information on contiguity is reported for more than half of the resources surveyed. At the same time, the combinatorial possibilities of MWEs are described in 3 lexica: one for Swedish (Borin et al., 2013), i.e., a FrameNet project, and two for Czech (Zabokrtský & Lopatková, 2007; Urešová et al., 2014).

Amaro et al. (2025) presents the results of a community survey on the usage of MWE resources (especially lexica) in various NLP tasks and downstream applications, showing that these resources can be appealing to developers due to their rich and diverse information.

Barbu Mititelu et al. (2025) describe the state of play in the development of MWE lexica in the years after Losnegaard et al. (2016)’s survey, focusing on the representation level of languages. Key trends identified include the predominance of monolingual resources developed for computational use, although bi- and multilingual resources are also reported; a shift toward language resources that are MWE-dedicated rather than MWE-aware; and the widespread inclusion of these resources in language technology catalogs.

Despite the interest in the way in which MWEs are dealt with in various language resources, none of these previous surveys have looked into the way in which their syntactic aspects are described in lexicographic works, which is our aim here.

4. Methodology and findings

Based on the list of MWE lexica from Barbu Mititelu et al. (2025), we further investigated which ones also address syntactic aspects of MWEs. We started from this list because the survey period covered (2016–2024) allows us to consider it up-to-date, and because the survey presents not only an extensive list of 66 resources but also supplies detailed notes on the lexical encoding of MWEs at various linguistic levels, including lemma information, syntactic behaviour, semantic representation, linking to corpora, etc. For our purposes, the syntactic information for each resource in Barbu Mititelu et al. (2025)’s list was collected in one of two ways: (i) we extracted information about the levels of description from the papers reporting on the resources; when no information was available, (ii) we accessed the resource (either online or after downloading the data provided) and investigated the description of MWEs.

Regarding the way MWEs are defined in each of these resources, we notice that this list of lexica covers various interpretations of the notion of MWEs, rather than a unified understanding of this linguistic phenomenon, including idioms, collocations, and expressions.

Four separate levels of syntactic information are considered: (i) morpho-syntactic information applying to the whole MWE, namely its PoS information which determines its paradigm – nominal, verbal, adverbial, etc.; (ii) the internal structure of the MWE represented as a linear sequence of PoS tags of its components, or as a tree structure; (iii) the variability of the syntactic realisation of the MWE in terms of word order, internal modification, discontinuity, alternations; and (iv) the subcategorisation or valency information of the MWE describing its compatibility with other phrases in text.

The main key findings can be summarised as follows:

- 53 out of 66 resources (80.3%) cover the morpho-syntactic description of MWEs; fewer resources, 41 out of the 66 (62.1%), are supplied with other levels of syntactic information.
- For 2 resources, the syntactic description is limited to alternations.
- 35 out of 53 monolingual resources are supplied with syntactic information, compared to 6 out of 13 bi- and multilingual resources.
- Only 7 resources are both supplied with syntactic information and linked to a corpus, and for only 3 resources the syntactic information is derived from a corpus.

Considering the type of the lexicon in terms of its purpose, i.e., exclusively covering MWEs or only including them among other lexical units, the analysis of the resources shows that only 3 out of 14 (21.4%) non-exclusive resources are supplied with any kind of syntactic information. It appears to be more typical for lexica dedicated exclusively to MWEs to offer this kind of information – 38 out of the 52 such lexica (73.1%).

5. Syntactic properties of MWEs in computational lexica: current trends

In this section, we discuss whether the lexica considered describe or encode several syntactic aspects of MWEs, and the way in which such information is described or encoded. However, due to space constraints, we will not provide examples here and encourage the reader to consult the respective papers for clarification.

5.1 Playing by the rules: MWE PoS

Information on the PoS of the MWEs is not always provided. This is the case of 13 out of the 66 lexica under scrutiny (those numbered [2], [5], [10], [14], [22], [24], [44], [45], [46], [48], [49], [51], [64], according to Table 4 in the Appendix). The remaining 53 resources present the information on the POS:

- Eighteen lexica cover only one PoS, as follows:
 - twelve contain only verbal MWEs ([21], [23], [25], [29], [32], [33], [34], [40], [47], [52], [54], [60]),
 - six contain only nominal MWEs ([4], [37], [39], [43], [55], [56]),
- both nominal and verbal MWEs are covered by six lexica ([26], [28], [30], [31], [41], [63]),
- two resources ([38] and [42]) cover nominal, verbal, and also functional MWEs,
- eight lexica cover nominal, verbal, and adjectival MWEs ([9], [36], [50], [59], [61], [62], [65], [66]),
- one resource also covers three PoS, this time nominal, adjectival, and adverbial ([11]),
- seventeen lexica cover all major PoS – nominal, verbal, adjectival and adverbial type of MWEs ([3], [6], [7], [8], [12], [13], [15], [16], [17], [18], [19], [20], [27], [35], [53], [57], [58]),

Type of MWE	Number	Percentage (%)
verbal	46	86.79
nominal	41	77.35
adjectival	27	50.94
adverbial	19	35.85
functional	4	7.55

Table 1: Distribution of MWEs of various PoSes in the 53 lexica

- one resource ([1]) covers all major PoS, as well as functional MWEs, which makes it the most comprehensive in this respect.

Table 1 illustrates the distribution of resources per PoS, based on the 53 resources:

Among the 18 lexica that cover only one PoS, the majority (two-thirds) focus on verbal MWEs (12 resources), while the remaining 6 target nominal MWEs. This suggests a stronger tendency to document verbal types when a single PoS is represented. In contrast, adverbial and functional MWEs appear only in resources that also include one or more major PoS types (nominal, adjectival, or verbal MWEs).

5.2 Internal structure

The information on the internal structure of MWEs appears in only half of the resources (51.5%)(Table 2).

Type of representation	Number	Percentage (%)
No representation	32	48.5%
PoS sequence (non-hierarchical)	21	31.8%
Partial tree structure (head only)	4	6.1%
Tree structure	9	13.6%

Table 2: Distribution of MWE lexica with respect to the representation of the internal structure of MWEs

A limited number of resources ([1], [29], [32], [38], [44], [52], [54], [58], [59]) represent the internal structure as a tree. Other resources specify only the head of the phrase, thus providing only a partial tree structure ([18], [20], [21], [50]). A larger proportion of the resources (31.8%) represents the MWE internal structure as a linear sequence of the PoS tags of its components. This type of syntactic description of the internal structure is often used to define further levels of description, such as modifications and word order permutations.

5.3 Argument structure

The argument structure of a MWE refers to the representation of its semantico-syntactic arguments. Such information is absent from 51 (i.e., 78%) of the 66 analyzed lexica. For the other 15 resources (i.e., 23%), such information is presented in various forms: 3 lexica ([24], [56], [57]) contain it in an underspecified way, i.e. by linkage to a corpus; one lexicon ([63]) only mentions the selected preposition categorized for by the MWEs; one resource maps each expression to a super-class representing a basic syntactic structure ([9]). The other 4 resources (6%) ([3], [38], [40], [53]) contain only the open slots that are usually rendered using indefinite pronouns or adverbs.

Only 6 resources in our pool of lexica encode the arguments of MWEs using various formalisms: two of them ([28], [46]) use lexical functions for rendering the combinatorial possibilities; one ([29]) merely takes such information over from the data sources (Wiktionary or wordnet examples); similarly, one lexicon ([59]) uses the notion of *frame* (as in FrameNet) to describe the arguments a MWE has; *catene* is the formalism used by another paper ([58]) at several linguistic levels, syntax included, for the description of the lexical entries; the Bulgarian-Romanian bilingual lexicon ([54]) specifies the arguments in terms of Universal Dependencies¹ syntactic relations. As the entries in this lexicon are translation equivalents (given the procedure the authors adopted in designing the lexicon, i.e., starting from equivalent wordnet synsets), such a resource offers insights into the syntactic behaviour in terms of combinatorial possibilities of equivalent MWEs.

Interestingly, all these 10 resources (the 4 ones containing indefinite pronouns and adverbs for open slots and the 6 ones that encode the argument structure of the contained MWEs) show language diversity, in the sense that they cover 11 languages: 8 lexica are monolingual, describing MWEs in Portuguese ([3]), Finnish ([29]), Arabic ([38]), Hebrew ([40]), Spanish ([46]), Dutch ([53]), Bulgarian ([58]), Greek ([59]). Two lexica are bilingual: one ([28]) includes French and English, and the other one ([54]) Bulgarian and Romanian. This indicates that specialists working with various languages have recognised the necessity to describe the combinatorial possibilities of MWEs in their respective languages.

However, interest in this syntactic aspect manifested itself both for resources that contain only verbal MWEs ([29], [54]), and for those that contain MWEs from other PoSes ([3], [28], [38], [40], [53], [58], [59]).

5.4 Word order variation

Some MWEs allow their components to change order while preserving meaning, which means that the status of the MWE remains unaffected. In such cases, nothing new is inserted, and only the sequence of fixed components changes. The easiest example is a reversible coordination of two adjacent words (e.g. “on and off” can be used as “off and on”). Still, similar rotations also occur in light-verb constructions when the noun and verb change positions in passive voice (e.g. “take charge of sth.” may be substituted with “charge was taken of sth.”). In our 66 resources, only 15 lexica (22.7%) (see Table 3) explicitly record at least one such permutation. Approximately 50 resources either suppose a single possible ordering or leave the issue unspecified.

¹ <https://universaldependencies.org/>

Strategies for covering word order variation are subdivided into two main groups: 1) those resources that have special field, tag or other label to address variation ([1], [26], [32], [53] etc.), and: 2) those resources in which permutations are visible only through linked corpora ([9], [24] and [57]). Notably, two-thirds of the resources with word order variation describe languages with free constituent order (e.g., Czech, Finnish, etc.), where permutation is common and essential. Additionally, it is worth noting that resources published after 2022 more frequently mention word order variation, whereas most earlier ones do not. To summarise, although word order permutations are essential for accurate parsing, generation, and cross-linguistic alignment, they remain underrepresented in the majority of current lexical resources.

Feature	Resources with support	Percentage (%)
Word-order permutations	15	22.7
Internal modification	13	19.7
Discontinuity	13	19.7
Verb alternations	14	21.2

Table 3: Coverage of four syntactic phenomena in the 66 lexical resources surveyed

5.5 Internal modification of MWEs

MWEs that allow for internal modification are idioms whose meaning remains intact even when (at least) one of their components is modified by (at least) a word that is not a component of the respective MWE. For lexicographers, the key task is to specify which MWEs license such ‘internal slots’ and what kinds of modifiers (adjectives, adverbs, numerals, etc.) may fill them. Across the 66 resources, only 13 lexica (approx. 19.7 %) (see Table 3) explicitly encode internal modification. Eleven of the lexica include a separate information field that notes MWEs with internal modifiers (e.g., [1], [3], [5], [38], among others), whereas two resources capture internal modification only through corpus data ([24], [59]). The remaining resources, 53 lexica, give no information on the matter. Notably, most of the lexica that capture internal modification focus on morphologically rich languages, where modifier insertion is both common and linguistically possible.

5.6 Representing discontinuities

Discontinuity is a syntactic property of an expression whose components do not form an uninterrupted string, i.e., words that are not components of a MWE can occur between its lexical elements. In the case of MWEs, discontinuity is essential for modeling the lexicon-grammar interface, as it requires specifying the constraints that govern gaps or where ellipsis can occur. Encoding discontinuity helps researchers identify correlations between word order flexibility, morphology, and idiomaticity. While discontinuity relates to internal modification as modifiers may appear between the fixed components of the MWE (see section 5.5), we consider it separately and examine whether the resources explicitly address discontinuity in the MWE structure.

An important issue with respect to discontinuity is the challenge it poses on parsing and tagging, since most traditional parsers are designed to recognise contiguous phrases and often struggle with multiword expressions (MWEs), because discontinuous or non-projective constructions usually involve long-distance dependencies (Johnson, 2002; Campbell, 2004; Nivre & Nilsson, 2005). Some specialised algorithms have been developed to handle discontinuity, e.g. in MaltParser (Nivre et al., 2006). More recent methods have also been explored which are able to handle relations between non-adjacent components, e.g. graph convolutional neural networks and attention-based methods (Rohanian et al., 2019), discontinuous constituency parsing by averaging the predicted trees (Shayegh et al., 2024), MWEasWSD approach combining a rule-based pipeline and a trainable bi-encoder model using also discontinuous training data (Tanner & Hoffman, 2023), etc. Ide et al. (2025) report that fine-tuning of LLMs outperform state-of-the-art in parsing MWEs, in particular discontinuous ones. While recent methods show promising results, MWE identification across discontinuities still remains a challenge, most notably with respect to the low recall (Ide et al., 2025), especially for unseen MWEs and low-resource languages, and the precise structural identification of the MWEs influence significantly their semantic analysis (Tayyar Madabushi et al., 2021; Miletic & Walde, 2024).

When we examine how MWE lexica encode or deal with discontinuity, the possibility that other words can split apart the pieces of a MWE yields inconsistent results. Across 66 lexica in our survey, 53 of them (approx. 80%) (see Table 3) do not provide any information about whether MWEs can break apart, and only 13 resources (approx. 20%) encode discontinuity in any form. These resources cover 11 languages, including Arabic, Czech, Greek, Hebrew, Indian languages, etc.

The resources can be subdivided into those that assign special features, tagsets, or substructure information that indicates whether a component can move and under what constraints. For example, one lexicon ([1]) uses a slot-based template, while some ([32], [54]) rely on a separation between fixed and variable segments. There are resources ([24], [53], [57], [59]) that do not include the information mentioned above, but point entries to a corpus, where information about such gaps can be deduced from empirical corpus data.

5.7 Verb alternations

Verb alternations refer to patterns where a verb can occur in different inflected forms while preserving its meaning; such alternations help emphasize the most important element of the sentence. Examples of alternations include active – passive constructions (Mary broke the vase. – The vase was broken by Mary.), causative alternation (the ice cube melted. – the sun melted the ice cube.), etc. (Levin, 1993). Depending on their internal arguments, argument roles, or licensing, many MWEs and, primarily, idioms do not readily provide verb alternations. Upon examining the data, it becomes apparent that similar issues affect the majority of the lexica. Only 14 lexica (21%) out of the total 66 (see Table 3) (corresponding to 30% of the 46 lexica containing verbal MWEs) include information about verbal alternations in MWEs. They describe 10 languages (Arabic, Czech, Hebrew, etc.). In these cases, verbal alternations are mainly represented by active-passive ([3], [23], [25], [38], [52], etc.) and rarely by dative shift ([59]) alternations. But the main challenge is that some lexica include entries with verbal alternations, but do not provide any special labeling for them. In contrast, others offer their complete annotations ([1], [25], and [59])

or the annotation of their morphological inflection, but not deep syntactic alternations ([9]).

6. Relevance of Syntactic Description of MWEs for NLP Tasks

In this section, we examine the relevance and usefulness of syntactic information encoded in lexical resources for various NLP tasks.

Previous literature has already highlighted the importance of a syntactic description of MWEs in lexica dedicated to them or containing them alongside other lexemes. Mohamed et al. (2024) suggests that lexica containing even basic morphosyntactic information, such as PoS, can enhance the annotated datasets and improve MWE identification strategies. Muller et al. (2024) demonstrates that lexica with more detailed syntactic description (such as the potential modifiers of the MWE as a whole or of its components) can make parsing MWEs more accurate. Recently, presenting the results of a community survey Amaro et al. (2025) reported that the respondents considered that the existing resources containing MWEs can be enhanced to increase their usefulness in NLP tasks by enriching them with linguistic information, more detailed and finer-grained syntactic description being one crucial type of such information, alongside more and better examples, as well as links to annotated corpora. The authors concluded that syntactic information contributes to modelling the behaviour of MWEs in corpora straightforwardly and explicitly.

In what follows, we present two specialised toolkits designed for identifying MWEs in corpora, outlining their features and their contribution to other NLP tasks. We discuss the relevant syntactic features in MWE identification with a view to developing MWE lexica with specific syntactic description of MWEs.

6.1 MWE Toolkit

The MWE Toolkit (Ramisch, 2015) is an open, language-independent framework for automatic MWE acquisition and identification in corpora. It is designed to discover, validate, and extract MWEs using user-specified syntactic patterns and statistical measures. The toolkit’s pipeline is modular: it first tags and lemmatizes a corpus, then applies morpho-syntactic extraction patterns, and finally scores candidate n-grams with association measures. In effect, it builds a lexical resource of MWEs annotated with their syntactic properties (e.g., PoS sequence) and frequency counts. The toolkit supports any language given an adequate PoS tagger, and has been applied successfully to model the lexical and syntactic behaviour of MWEs in many languages (Cordeiro et al., 2016; Scholivet et al., 2018).

The MWE Toolkit leverages syntactic information using morpho-syntactic patterns, which are user-defined sequences of PoS tags (e.g., NOUN + ADJ, VERB + ADP + NOUN). These patterns are applied to a lemmatized and PoS-tagged corpus to extract candidate MWEs. Each extracted candidate is stored with its lemmas, PoS sequence, and frequency/association scores. This representation captures the syntactic ‘shape’ of an expression, allowing downstream systems to treat it as a single structured unit (Zagatti et al., 2022).

The Toolkit focuses on basic grammatical features like PoS tags and word order instead of full dependency parses. However, these simple features prove effective, as MWEs often

follow rigid syntactic templates (e.g., the MWE “kick the bucket” matches VERB + DET + NOUN). Handling these patterns filters out irrelevant n-grams and highlights linguistically plausible MWEs (Ramisch et al., 2010). This approach helps limit false positives and improves the precision of the Toolkit. It also enables easier integration with other tools; for instance, parsers can treat matched MWEs as atomic units, reducing parsing errors. By providing transparent and language-independent syntactic filters, the MWE Toolkit enables the direct usability of structured MWE knowledge for various NLP tasks.

Originally developed around 2010 (Ramisch et al., 2010), the Toolkit has undergone continuous development since its inception. Initially, before 2015, it was based primarily on surface forms, lemmas, PoS tags, and PoS patterns for the generation and filtering of candidates for MWEs, with the acknowledged but unintegrated understanding that deeper syntactic information could improve results (Ramisch, 2015). By 2016, with ‘mwe-toolkit+sem’, the toolkit began to capture deeper syntactic characteristics through word embeddings implicitly, utilising PoS tags and lemmas as input for the semantic compositionality scoring based on cosine distance between MWE vectors and their component words (Cordeiro et al., 2016). The evolution continued into 2022 with ‘mwetoolkit-lib’, which directly uses PoS tags and lemmas within its MWE extraction pipeline, but crucially facilitates integration with external parsing pipelines and sophisticated NLP tools, allowing for MWE-based feature enrichment, such as document clustering (Zagatti et al., 2022).

The accurate identification and processing of MWEs are crucial for numerous downstream NLP applications. The Toolkit, through its core functionalities and subsequent extensions, has demonstrated its utility in enhancing the performance and linguistic accuracy of various tasks:

- **Parsing:** Identifying MWEs as single units can noticeably improve parsing accuracy. In practice, one common strategy is to concatenate recognised MWEs before parsing, turning “prime minister” into one token, for example. Scholivet et al. (2018) note that this simple pre-processing (treating MWEs as “words with spaces”) is an effective way to reduce attachment errors. By embedding the Toolkit’s MWE lexicon into a parser’s pipeline, parsers encounter fewer ambiguous constructions (e.g., a fixed verb-preposition will not attract a random object).
- **Machine Translation (MT):** Treating MWEs as units often leads to more fluent translations, especially in statistical and neural MT. Riktors & Bojar (2017) used the MWE Toolkit to extract c. 400,000 English–Czech and 60,000 English–Latvian MWE pairs from parallel corpora and integrated them into the training of neural MT systems yielding a modest BLEU improvement of 0.99 on an MWE-specific test data set. Similarly, Ebrahim et al. (2017) proposed a new method for detecting and integrating English phrasal verbs as a type of MWEs into an English–Arabic phrase-based statistical MT system. The MWE detection process specifically utilised the MWE-toolkit. While the baseline system achieved the highest BLEU score, the study suggests that detecting more linguistic patterns and integrating them into the En-Ar statistical MT system could improve translation quality with other integration methods, highlighting that linguistic features are not appropriately handled in statistically learned models.
- **Information Extraction (IE):** Syntactic MWE data can also aid IE tasks, such as event extraction and relation detection. For example, Sánchez Cárdenas & Ramisch (2019) used the MWE Toolkit in combination with lexicon- and pattern-based

methods to extract predicate frames from environmental science corpora. They automatically extracted verb-noun-noun triples (e.g., forests destroy resources) by bootstrapping noun-verb-noun MWE patterns, leveraging the syntactic extraction capabilities of the Toolkit. This study demonstrates how structured MWE representations, specifically those enriched with syntactic information, can populate semantic frames and support downstream IE tasks that require accurate identification of multi-word predicates and their arguments.

- **Other tasks (sentiment analysis, summarization, etc.):** Although few studies explicitly evaluate the MWE Toolkit in the context of sentiment analysis or text summarization, the underlying principle remains clear: many expressions that convey sentiment or key content are multiword. Idiomatic phrases like “kick the bucket” or “bright future” can significantly alter the tone or meaning of a sentence. Leveraging a lexicon of MWEs ensures that such expressions are correctly recognised and interpreted. Work by Duran & Ramisch (2011) has shown that lexical-syntactic patterns are effective in capturing sentiment-bearing expressions, supporting the idea that syntactic MWE patterns can enhance sentiment detection. It stands to reason that modern NLP systems could apply the pattern-based approach of the MWE Toolkit to improve sentiment classification or preserve the meaning of MWEs during summarization – although the targeted evaluation in these domains remains limited.

6.2 MWE-Finder: In search of Non-Adjacent and Flexible MWEs

Another notable tool that leverages the syntactic and morphological properties of MWEs that are encoded in a computational lexicon is MWE-Finder (Odijk et al., 2024). It is designed to search for MWEs in large, dependency-annotated corpora, with a particular focus on flexible expressions, i.e., those whose components can appear in different forms, orders, and at varying distances from each other.

There are two key-features of MWE-finder: (i) on the one hand, the system uses information encoded in the DUTch CAnonicalised Multiword Expressions lexical resource (DUCAME), a sophisticated and meticulously curated computational lexicon that contains more than 10,000 MWEs for the Dutch language in a canonical form; and on the other hand, (ii) the system is integrated into GrETEL,² a Dutch user-friendly linguistic search tool designed to explore treebanks (Augustinus et al., 2017). Users can either input a canonical form of a MWE or select from a list of about 10,000 predefined Dutch MWEs. This canonical form acts as a hypothesis about the MWE’s properties. Unlike standard search tools, MWE-Finder considers the grammatical structure of MWEs. It should be noted that the tool generates appropriate search queries automatically, finds matching sentences, detects unexpected modifiers or determiners, and provides detailed statistics on the syntactic behaviour of any given MWE and its components. MWE-Finder is designed for linguists and lexicographers researching MWEs, with a particular emphasis on flexible ones. While the current implementation of the tool handles MWEs in Dutch, the system is also portable to other languages.

² Greedy Extraction of Trees for Empirical Linguistics

6.3 Syntactic information used in MWE identification

The PARSEME shared tasks on the identification and classification of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018, 2020) have provided a controlled environment for testing various methods and techniques for identifying MWEs in text. These shared tasks offer the opportunity to address specific aspects of the automatic treatment of MWEs, such as word order variations, discontinuity, and distinguishing between literal and idiomatic meanings, among others.

It has often been noted that MWE lexicons are essential for identifying MWEs. In particular, Savary et al. (2019) emphasizes the need to couple identification with MWE discovery via syntactic MWE lexicons, which should provide minimal morphosyntactic information to complement existing MWE-annotated corpora.

We have analysed 25 papers reporting on systems or methods that participated in the PARSEME shared tasks or were presented at various conferences, e.g., the MWE workshop,³ LREC,⁴ etc.

These studies demonstrate that the identification of MWEs based on syntactic properties derived from a lexicon is rare and mostly uses dependency information (Bejček et al., 2013). Other data-driven systems rely on syntactic information derived through corpus annotation (PoS tagging) and parsing without prior lexical description of the syntactic properties of MWEs, e.g., the system TRAVERSE (Waszczuk, 2018), ranked first at PARSEME shared task in 2018; Boros et al. (2017) reporting various evaluation scores for different languages; the system HMSid (Colson, 2020) using syntactic patterns and association measures. Moreau et al. (2018) compares two systems — one based on sequential analysis and the other on dependency trees. The authors demonstrate that dependency-based annotation enhances the identification, particularly of discontinuous MWEs.

One of the main syntactic features of MWEs is their syntactic fixedness – expressed as fixed word order, restrictions on possible modifications, restrictions on alternations, etc. These are usually encoded in dictionaries, but they have been rarely employed in MWE identification systems, as such dictionaries are rare or the syntactic description is not formalized. This is one possible extension of the syntactic description of MWE lexica, which can boost MWE identification.

7. Conclusions and outlook for further research

In this paper, we present the current state-of-the-art in MWE lexicon development, with a particular emphasis on the encoding of syntactic properties of MWEs.

Ultimately, despite the remarkable capabilities of LLMs in generating fluent and contextually appropriate language, they often struggle with the systematic treatment of idiosyncratic, discontinuous, or low-frequency MWEs, especially in less-resourced languages or specialized domains (Tayyar Madabushi et al., 2021; Chakrabarty et al., 2022; Phelps et al., 2024; Miletić & Walde, 2024). These expressions frequently fall “below the surface” of purely distributional learning. A syntax-oriented survey of MWE lexica is therefore timely and relevant, as it provides structured insights into how MWEs can be

³ <https://multiword.org/events/previousevents>

⁴ <https://www.elra.info/elra-events/lrec/>

explicitly modeled and anchored in interpretable, symbolic resources. Such lexica not only support error analysis and fine-tuning of LLMs, but can also serve as grounding tools in hybrid systems, Retrieval-Augmented Generation (RAG) architectures, or evaluation benchmarks that aim to improve MWE understanding in current and future language models.

8. Acknowledgments

This work received support from the CA21167 COST action Universality, diversity and idiosyncrasy in language technology (UniDive), funded by COST (European Cooperation in Science and Technology). Also, part of this research was supported by Aristotle University of Thessaloniki (Grant YP3TA-0561062) and the Ministry of Science, Republic of Serbia #GRANT 451-03-136/2025-03/200174. Another part of this research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

9. References

- Amaro, R., Giouli, V., Korvel, G., Lobzhanidze, I., Barbu Mititelu, V. & Valunaite Oleskeviciene, G. (2025). Perceptions on MWE lexicons use in NLP by the User Community: features, challenges and recommendations. URL https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:27_perceptions_on_mwe_lexicons.pdf. 3rd UniDive Workshop in Budapest.
- Augustinus, L., Vandeghinste, V., Schuurman, I. & Van Eynde, F. (2013). Example-Based Treebank Querying with GrETEL—Now Also for Spoken Dutch. In S. Oepen, K. Hagen & J.B. Johannessen (eds.) *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Oslo, Norway: Linköping University Electronic Press, Sweden, pp. 423–428. URL <https://aclanthology.org/W13-5638/>.
- Augustinus, L., Vandeghinste, V., Schuurman, I. & Van Eynde, F. (2017). GrETEL: A Tool for Example-Based Treebank Mining. In J. Odiijk & A. van Hessen (eds.) *CLARIN in the Low Countries*, chapter 22. London, UK: Ubiquity, pp. 269–280. License: CC-BY 4.0.
- Baldwin, T. & Kim, S.N. (2010). Multiword Expressions. In F.J. Damerau & N. Indurkha (eds.) *Handbook of Natural Language Processing*. Chapman and Hall/CRC, 2 edition, pp. 267–292.
- Barbu Mititelu, V., Giouli, V., Korvel, G., Liebeskind, C., Lobzhanidze, I., Makhachashvili, R., Markantonatou, S., Markovic, A. & Stoyanova, I. (2025). Survey on Lexical Resources Focused on Multiword Expressions for the Purposes of NLP. In A.K. Ojha, V. Giouli, V.B. Mititelu, M. Constant, G. Korvel, A.S. Doğruöz & A. Rademaker (eds.) *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*. Albuquerque, New Mexico, U.S.A.: Association for Computational Linguistics, pp. 41–57. URL <https://aclanthology.org/2025.mwe-1.6/>.
- Bejček, E., Straňák, P. & Pecina, P. (2013). Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. In V. Kordoni, C. Ramisch & A. Villavicencio (eds.) *Proceedings of the 9th Workshop*

- on *Multiword Expressions*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 106–115. URL <https://aclanthology.org/W13-1016/>.
- Borin, L., Forsberg, M. & Lyngfelt, B. (2013). Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. *Veredas*, 17(1), pp. 28–43.
- Boros, T., Pipa, S., Barbu Mititelu, V. & Tufis, D. (2017). A data-driven approach to verbal multiword expression detection. PARSEME Shared Task system description paper. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain: Association for Computational Linguistics, pp. 121–126. URL <https://aclanthology.org/W17-1716/>.
- Campbell, R. (2004). Using linguistic principles to recover empty categories. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Chakrabarty, T., Saakyan, A., Ghosh, D. & Muresan, S. (2022). FLUTE: Figurative Language Understanding through Textual Explanations. In Y. Goldberg, Z. Kozareva & Y. Zhang (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7139–7159. URL <https://aclanthology.org/2022.emnlp-main.481/>.
- Colson, J.P. (2020). HMSid and HMSid2 at PARSEME Shared Task 2020: Computational Corpus Linguistics and unseen-in-training MWEs. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova & A. Savary (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online: Association for Computational Linguistics, pp. 119–123. URL <https://aclanthology.org/2020.mwe-1.15/>.
- Cook, P., Fazly, A. & Stevenson, S. (2007). Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context. In N. Gregoire, S. Evert & S.N. Kim (eds.) *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Prague, Czech Republic: Association for Computational Linguistics, pp. 41–48. URL <https://aclanthology.org/W07-1106/>.
- Cordeiro, S., Ramisch, C. & Villavicencio, A. (2016). mwetoolkit+ sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. pp. 1221–1225.
- Duran, M.S. & Ramisch, C. (2011). How do you feel? investigating lexical-syntactic patterns in sentiment expression. In *Proceedings of corpus linguistics*.
- Ebrahim, S., Hegazy, D., Mostafa, M.G.H.M. & El-Beltagy, S.R. (2017). Detecting and integrating multiword expression into English-Arabic statistical machine translation. *Procedia Computer Science*, 117, pp. 111–118.
- Giouli, V. & Barbu Mititelu, V. (2024). *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Language Science Press.
- Ide, Y., Tanner, J., Nohejl, A., Hoffman, J., Vasselli, J., Kamigaito, H. & Watanabe, T. (2025). CoAM: Corpus of All-Type Multiword Expressions. In W. Che, J. Nabende, E. Shutova & M.T. Pilehvar (eds.) *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 27004–27021. URL <https://aclanthology.org/2025.acl-long.1311/>.
- Iñurrieta, U., Aduriz, I., Díaz de Ilarraza, A., Labaka, G. & Sarasola, K. (2018). Konbitzul: an MWE-specific database for Spanish-Basque. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.) *Proceedings of the Eleventh*

- International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1397/>.
- Johnson, M. (2002). A simple pattern-matching algorithm for recovering empty nodes and their antecedents. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press.
- Lion-Bouton, A., Savary, A. & Antoine, J.Y. (2023). A MWE lexicon formalism optimised for observational adequacy. In A. Bhatia, K. Evang, M. Garcia, V. Giouli, L. Han & S. Taslimipour (eds.) *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 121–130. URL <https://aclanthology.org/2023.mwe-1.16/>.
- Losnegaard, G.S., Sangati, F., Escartín, C.P., Savary, A., Bargmann, S. & Monti, J. (2016). PARSEME Survey on MWE Resources. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2299–2306. URL <https://aclanthology.org/L16-1364>.
- Maziarz, M., Grabowski, Ł., Piotrowski, T., Rudnicka, E. & Piasecki, M. (2023). Lexicalisation of Polish and English word combinations: an empirical study. *Poznan Studies in Contemporary Linguistics*, 59(2), pp. 381–406.
- Miletić, F. & Walde, S.S.i. (2024). Semantics of Multiword Expressions in Transformer-Based Models: A Survey. *Transactions of the Association for Computational Linguistics*, 12, pp. 593–612. URL <https://aclanthology.org/2024.tacl-1.33/>.
- Mohamed, N.H., Savary, A., Khelil, C.B., Antoine, J.Y., Keskes, I. & Belguith, L.H. (2024). Lexicons Gain the Upper Hand in Arabic MWE Identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024*. pp. 88–97.
- Moreau, E., Alsulaimani, A., Maldonado, A. & Vogel, C. (2018). CRF-Seq and CRF-DepTree at PARSEME Shared Task 2018: Detecting Verbal MWEs using Sequential and Dependency-Based Approaches. In A. Savary, C. Ramisch, J.D. Hwang, N. Schneider, M. Andresen, S. Pradhan & M.R.L. Petruck (eds.) *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 241–247. URL <https://aclanthology.org/W18-4926/>.
- Muller, I., Mamede, N. & Baptista, J. (2024). Hurdles in Parsing Multi-word Adverbs: Examples from Portuguese. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H.G. Oliveira & R. Amaro (eds.) *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, pp. 177–186. URL <https://aclanthology.org/2024.propor-1.18>.
- Nivre, J., Hall, J. & Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).
- Nivre, J. & Nilsson, J. (2005). Pseudo-Projective Dependency Parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 99–106.

- Odiijk, J., Kroon, M., Spoel, S., Bonfil, B. & Baarda, T. (2024). *Querying for multiword expressions in large Dutch text corpora*. *Phraseology and Multiword Expressions*. Language Science Press, pp. 229–267. Publisher Copyright: © 2024, the authors. All rights reserved.
- Phelps, D., Pickard, T.M.R., Mi, M., Gow-Smith, E. & Villavicencio, A. (2024). Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In A. Bhatia, G. Bouma, A.S. Doğruöz, K. Evang, M. Garcia, V. Giouli, L. Han, J. Nivre & A. Rademaker (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL, pp. 178–187. URL <https://aclanthology.org/2024.mwe-1.22>.
- Ramisch, C. (2015). Multiword Expressions Acquisition: A Generic and Open Framework.
- Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A. & Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In A. Savary, C. Ramisch, J.D. Hwang, N. Schneider, M. Andresen, S. Pradhan & M.R.L. Petruck (eds.) *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 222–240. URL <https://aclanthology.org/W18-4925/>.
- Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A. & Xu, H. (2020). Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova & A. Savary (eds.) *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online: Association for Computational Linguistics, pp. 107–118. URL <https://aclanthology.org/2020.mwe-1.14/>.
- Ramisch, C., Villavicencio, A. & Boitet, C. (2010). Multiword expressions in the wild? the mwetoolkit comes in handy. In *Coling 2010: Demonstrations*. pp. 57–60.
- Rikters, M. & Bojar, O. (2017). Paying Attention to Multi-Word Expressions in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVI: Research Track*. pp. 86–95.
- Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L.A. & Mitkov, R. (2019). Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In J. Burstein, C. Doran & T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2692–2698. URL <https://aclanthology.org/N19-1275/>.
- Rosen, V., Losnegaard, G.S., De Smedt, K., Bejcek, E., Savary, A., Przepiorkowski, A., Osenova, P. & Barbu Mititelu, V. (2015). A survey of multiword expressions in treebanks. In *Proceedings of the Treebanks and Linguistic Theories conference (TLT - 2015)*. Warsaw, Poland, pp. 179–193. URL <https://ufal.mff.cuni.cz/biblio/attachments/2015-bejcek-m9168853603341436530.pdf>.
- Sag, Ivan A. and Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276. Springer, pp. 189–206.

- Sánchez Cárdenas, B. & Ramisch, C. (2019). Eliciting specialized frames from corpora using argument-structure extraction techniques. *Terminology*, 25(1), pp. 1–31.
- Savary, A., Cordeiro, S. & Ramisch, C. (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)*. Florence, pp. 79–91.
- Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I. & Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain: Association for Computational Linguistics, pp. 31–47. URL <https://aclanthology.org/W17-1704/>.
- Scholivet, M., Ramisch, C. & Cordeiro, S. (2018). Sequence models and lexical resources for MWE identification in French. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 2. Language Science Press, p. 263.
- Shayegh, B., Wen, Y. & Mou, L. (2024). Tree-Averaging Algorithms for Ensemble-Based Unsupervised Discontinuous Constituency Parsing. In L.W. Ku, A. Martins & V. Srikumar (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 15135–15156. URL <https://aclanthology.org/2024.acl-long.808/>.
- Skoumalová, H. & Kopřivová, M. (2024). *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, chapter LEMUR: A lexicon of Czech multiword expressions. Language Science Press.
- Tanner, J. & Hoffman, J. (2023). MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation. In H. Bouamor, J. Pino & K. Bali (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 181–193. URL <https://aclanthology.org/2023.findings-emnlp.14/>.
- Tayyar Madabushi, H., Gow-Smith, E., Scarton, C. & Villavicencio, A. (2021). AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models. In M.F. Moens, X. Huang, L. Specia & S.W.t. Yih (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3464–3477. URL <https://aclanthology.org/2021.findings-emnlp.294/>.
- Urešová, Z., Štěpánek, J., Hajič, J., Panevova, J. & Mikulová, M. (2014). PDT-Vallex: Czech Valency lexicon linked to treebanks. URL <http://hdl.handle.net/11858/00-097C-0000-0023-4338-F>. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Vondříčka, P. (2019). Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics*, 112(1), pp. 83–101.
- Waszczuk, J. (2018). TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model. In A. Savary, C. Ramisch, J.D. Hwang, N. Schneider, M. Andresen, S. Pradhan & M.R.L. Petruck (eds.) *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 275–282. URL <https://aclanthology.org/W18-4931/>.

- Zabokrtský, Z. & Lopatková, M. (2007). Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. *Prague Bull. Math. Linguistics*, 87, pp. 41–60. URL <http://ufal.mff.cuni.cz/pbml/87/zabokrtsky-lopatkova.pdf>.
- Zagatti, F., de Lima Medeiros, P.A., da Cunha Soares, E., dos Santos Silva, L.N., Ramisch, C. & Real, L. (2022). mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022*. pp. 112–117.
- Zhang, X., Bouma, G. & Bos, J. (2025). Neural Semantic Parsing with Extremely Rich Symbolic Meaning Representations. *Computational Linguistics*, 51(1), pp. 235–274. URL https://doi.org/10.1162/coli_a_00542. https://direct.mit.edu/coli/article-pdf/51/1/235/2483888/coli_a_00542.pdf.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

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



Appendix

Table 4: List of resources from Barbu Mititelu et al. (2025) with the morphosyntactic types of MWEs they cover: N – Nominal MWEs; V – Verbal MWEs; Adj – Adjectival MWEs; Adv – Adverbial MWEs; Func – Functional MWEs.

#	Lexicon name and link	N	V	Adj	Adv	Func
1	LEMUR	✓	✓	✓	✓	✓
2	NileULex					
3	LEX-MWE-PT - Word Combinations in Portuguese	✓	✓	✓	✓	
4	Lexicalisation of Polish and English word combinations	✓				
5	The Database of Lithuanian MWEs					
6	Automatically constructed multiword lexicon srMWELEX v0.5	✓	✓	✓	✓	
7	hrMWELEX – Croatian lexicon of multiword expressions	✓	✓	✓	✓	
8	slMWELEX – Slovene lexicon of multiword expressions	✓	✓	✓	✓	
9	Srp DELAC	✓	✓	✓		
10	Expressions (deChile)					
11	Czech MWEs	✓		✓	✓	
12	The Dictionary of Estonian Phraseology	✓	✓	✓	✓	
13	Terminological MWE lexicon	✓	✓	✓	✓	
14	Terminology database of expressions					
15	Idioms of Chile [Chilenismos]	✓	✓	✓	✓	
16	Lunfardo Dictionary	✓	✓	✓	✓	
17	The Dictionary of Estonian Synonyms	✓	✓	✓	✓	
18	ilFhocail	✓	✓	✓	✓	
19	Referentiebestand Belgisch-Nederlands	✓	✓	✓	✓	
20	Czech Dependency Bigrams from the Prague Dependency Treebank	✓	✓	✓	✓	
21	Konbitzul		✓			
22	English-Persian database of idioms and expressions					
23	ParaDi 2.0 dataset - second version		✓			
24	MWEs lexicon extracted from the Gigafida 2.1 corpus					
25	Czech Verbal MWEs		✓			
26	Bulgarian and Romanian MWE dictionary	✓	✓			
27	ConceptNet-el	✓	✓	✓	✓	✓
28	CollFrEn: Rich Bilingual English–French Collocation Resource	✓	✓			
29	FinnMWE: a lexicon of Finnish MWEs		✓			

#	Lexicon name and link	N	V	Adj	Adv	Func
30	Russian Collocations Database	✓	✓			
31	Diretes (Diccionario RETicular de ESpañol)	✓	✓			
32	IDION: A database for Modern Greek MWEs		✓			
33	PolylexFLE		✓			
34	Japanese compound verb lexicon		✓			
35	Sentiment Lexicon of Idiomatic Expressions (SLIDE)	✓	✓	✓	✓	
36	LIDIOMS: A Multilingual Linked Idioms Data Set	✓	✓	✓		
37	LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds	✓				
38	SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal MWEs	✓	✓			✓
39	How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality	✓				
40	A Lexical Resource of Hebrew Verb-Noun MWEs		✓			
41	Towards Lexical Encoding of MWEs in Spanish Dialects	✓	✓			
42	MWE Dataset for Indian Languages	✓	✓			✓
43	Noun Compound Senses (NCS) dataset	✓				
44	MWE Dataset for Swedish					
45	Noun Compound Dataset for Russian					
46	Diccionario de Colocaciones del Español (DiCE)					
47	Polish verbal MWEs		✓			
48	Dutch idiomatic expressions					
49	MWE combinet 1.0					
50	Grammatical Dictionary of Multiword Units (Słownik gramatyczny jednostek wielowyrazowych)	✓	✓	✓		
51	Dictionary of idioms for Georgian					
52	IDION POMAK		✓			
53	DUCAME	✓	✓	✓	✓	
54	MWE dictionary for Bulgarian and Romanian		✓			
55	Feature-NN	✓				
56	Reddy-NN	✓				
57	MWE-CEFR Profiles	✓	✓	✓	✓	
58	Bulgarian Integrated Lexicon	✓	✓	✓	✓	
59	MWEs in FrameNet-EL	✓	✓	✓		
60	Verbal Multi-Word Expressions in Yiddish		✓			
61	IdiomKB	✓	✓	✓		
62	870 English idioms: norming and statistical analysis	✓	✓	✓		

#	Lexicon name and link	N	V	Adj	Adv	Func
63	Collocational Database for Learning Croatian as a Foreign Language	✓	✓			
64	Normed lexicon of English and Italian idioms					
65	Croatian dictionary of idioms	✓	✓	✓		
66	IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions	✓	✓	✓		