Chapter 1

Multiword Expressions lexica for Natural Language Processing: A state-of-the-art survey

Stella Markantonatou[a], Ivelina Stoyanova[b], Verginica Barbu Mititelu[c], Voula Giouli[d], Aleksandra Marković[e], Irina Lobzhanidze[f], Rusudan Makhachashvili[g], Gražina Korvel[h] & Chaya Liebeskind[i]

[a]Institute for Language and Speech Processing, ATHENA RC [b]Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences [c]Research Institute for Artificial Intelligence, Romanian Academy [d]Aristotle University of Thessaloniki [e]Institute for Serbian Language [f]Institute of Linguistic Studies, Ilia State University [g]Borys Grinchenko Kyiv Metropolitan University [h]Vilnius University [i]Jerusalem College of Technology

This chapter presents the results of a survey on lexica containing multiword expressions, which offers an updated perspective on this field of research as compared to the survey by Losnegaard et al. (2016), and also widens the discussion. We identified 66 resources, predominantly in the span of 2016-2024, both lexica dedicated to multiword expressions and general lexica featuring multiword expressions. To this end, we conducted a literature survey, as well as a survey sent out to the community. We analyzed these 66 resources from several perspectives: types of lexica, types of multiword expressions contained, languages covered, linkage to other resources, acquisition method, and representativeness. Further, we provide a more in-depth analysis of how the lemma of the multiword expressions is encoded in these resources and how their meaning is represented. In the context of the current trend of using large language models in language processing, we find strong indications that the linguistic information from such lexica can help enhance their performance in dealing with multiword expressions. The observations on the existing multiword expressions lexica motivate recommendations for some best practices in their development.



Stella Markantonatou, Ivelina Stoyanova, Verginica Barbu Mititelu, Voula Giouli, Alexandra Markovic, Irina Lobzhanidze, Rusudan Makhachashvili, Gražina Korvel & Chaya Liebeskind. 2026. Multiword Expressions lexica for Natural Language Processing: A state-of-the-art survey. In Verginica Barbu Mititelu & Voula Giouli (eds.), *Multiword expressions in Natural Language Processing: Current trends and challenges*, 1–67. Berlin: Language Science Press. DOI: 10.5281/zenodo.21277667

1 Introduction

Multiword expressions (MWEs) present significant challenges in Natural Language Processing (NLP), primarily due to their semantic non-compositionality, among other challenging characteristics. The expression 'pain in the neck' meaning "annoyance or source of grief", is a well-known example of a verbal MWE (VMWE). These characteristics render automatic identification of MWEs in text essential for semantically driven downstream applications. Despite recent advances, including the advent of large (and small) language models, the inherent complexity and distributional properties of MWEs continue to impede their effective automatic processing. Lexical resources, i.e., computational lexica dedicated to MWEs, are essential to address these challenges (Savary et al. 2019a).

In this chapter, we delve into the current state-of-the-art in MWE lexicon development, aiming to provide a comprehensive overview of the existing resources and the information they include about MWEs. We envisage this study as a basis for further work on questions such as "which available information about MWEs is useful to state-of-the-art NLP?" and "what are the best practices for encoding useful information about MWEs?". This way, the chapter will foster improvements in both the development and deployment of MWE lexica, to ensure that they can support future NLP innovations more effectively.

Furthermore, we propose a strategic roadmap for the MWE research community, emphasizing the need for more robust MWE-informed lexica that offer essential linguistic information. This roadmap primarily outlines a proposal for guidelines or minimum requirements for the development of MWE lexicon.

The chapter is structured as follows: in Section 2, we present the objectives and scope of this critical literature review; in Section 3, we describe the survey methodology, the repositories that were queried to identify the resources, and the organization of the spreadsheet containing the inventory of these resources and the type of data we include in it. Section 4 deals with the inclusion criteria and offers an overview of the lexical resources included in the review along several axes: (a) time span; (b) foreseen usage; (c) languages covered and linguality type; (d) types of lexica based on their content; (e) acquisition method; (f) accessibility; (g) representativeness, and (h) linking of MWE lexica to other resources. Section 5 contains a discussion of the form used to represent a MWE in a lexicon, or its lemma, as well as of several aspects involving the variation of a MWE and the way in which they are dealt with on the level of lemma representation. Section 6 discusses the way the meaning of MWEs is described and/or represented in lexica. Currently, with the advances Large Language Models (LLMs) have brought to the

NLP community (but not only), questioning the need for MWE lexica (just as that of language resources in general) has become legitimate. Section 7 deals with these aspects. Finally, Section 8 contains a discussion about the most important aspects addressed in the chapter and opens the way to defining the best practices in the development of MWE lexica. Finally, Section 9 wraps up the chapter with some concluding remarks about the inventory of MWE lexica and suggestions for further research. Limitations of our work are also presented.

2 Survey objectives

Our survey builds on the work of Losnegaard et al. (2016) (henceforth 'the PARSEME survey') that, in the framework of the PARSEME COST Action[1], provided a comprehensive overview of MWE resources, including lists, (MWE-aware and MWE-dedicated) lexica, and treebanks available at that moment. It was based on three language resources platforms: META-SHARE (Piperidis 2012), ELRA[2] and SIGLEX-MWE[3], in which the resources were found by keyword search. In addition, the linguistic community was asked to fill out a form about relevant resources they were aware of. General information about each resource was recorded: its name, a link to it, its type, contact information, the language(s), its size, the maximum length of the contained MWEs, whether it includes noncontinuous expressions, its license and accessibility policies, as well as some more advanced information: relevant publications describing it, its special MWE features and the grammatical or lexical formalism (when applicable).

Our work extends the scope of the PARSEME survey by exploring and updating the state of MWE resources from 2016 to date, paying extra attention to how the canonical form and the meaning of MWEs are represented, and addressing new developments and emerging gaps. This new survey is conducted within the UniDive COST Action (Savary et al. 2024) as an integral part of Working Group 2 that seeks to define the lexicon-corpus interface (Barbu Mititelu et al. 2024), focusing on specific aspects of the lexica that warrant further investigation. This aim follows a vision for creating an interconnected ecosystem of lexica and corpora, where MWEs are systematically represented and annotated, fostering more comprehensive and effective NLP solutions.

The primary objective of this survey is to provide a comprehensive overview of the current landscape of MWE-related computational lexica; the ongoing developments in the field of LLMs are also taken into account. The identification

[1]https://typo.uni-konstanz.de/parseme/
[2]https://www.elra.info/
[3]https://multiword.org/

of relevant resources was intended to be as exhaustive as possible. Therefore, we used strict inclusion criteria to ensure the relevance of the selected resources. Special emphasis is placed on the languages featured in the resources and the levels of representation of these languages. One of the main pillars of the UniDive Action is to foster awareness and understanding of language diversity and inclusion in Language Technology; by gathering information on the languages represented in MWE lexica, this survey aims to serve as a first step in highlighting the extent to which less-represented languages are included and supported in existing resources.

Beyond merely cataloging MWE lexica, our survey seeks to investigate the strategies developers use to encode lexical information for MWEs, particularly focusing on how the idiosyncrasies of these complex language constructs are represented. In this regard, the main goal of the survey is to delve into the lexical encoding these resources provide at various levels of analysis, including the 'canonical form' ('lemma of the MWE' is used as an equivalent term in this survey), syntactic behavior, and semantic representation of the MWE, its linking to corpora, etc. However, we do not discuss the encoding formats used by the resources, as this goes beyond the scope of this study.[4]

Lastly, after outlining the picture of the current MWE resource landscape, this survey aims to discuss whether the described situation aligns with the current and anticipated needs of the NLP stakeholders.

By addressing these objectives, we intend to reinforce ongoing work on the construction and enhancement of MWE lexica, while also providing targeted recommendations to inform their future development. The ultimate goal is to leverage these resources for the automatic identification of MWEs in text, their analysis and linking with corpora, thus enhancing their utility in NLP applications.

3 Methodology

We aimed to create a comprehensive collection of relevant data that would enable us to depict an accurate picture of the MWE resource landscape by cataloging MWE resources and detailing their properties. To achieve this, we defined the criteria for resource inclusion, which focused on retaining only computational lexica, databases, and lists centred on MWEs while excluding corpora, terminological databases, and named entity lists, thus departing from the approach of

[4]For a detailed discussion on the encoding format topic, see Lichte et al. (2019).

Losnegaard et al. (2016) which considered both lexical resources and parsed corpora, i.e., treebanks, in the survey.

A collaborative spreadsheet was created to systematically record, organize, and describe the collected resources. In the rest of this text we expand on the information documented in the spreadsheet. A detailed description of the spreadsheet is given in Section 3.2.

3.1 Sources

The sources for collecting information about MWE lexica include the following major repositories and databases:

1. *European Language Grid (ELG)*[5] (Rehm 2023). This is the largest platform where language technologies and language resources alike, developed by public or private bodies, are cataloged (and some of them even stored) to make them known to potential users and other developers and to facilitate access to them. The catalogue can be searched with keywords. To find the lexical resources of MWEs, we searched within the category *Lexical / Conceptual Resource* using the word 'expressions' and obtained 71 results. We have examined their description to decide upon their inclusion in the dataset.

2. *ACL Anthology*[6]. This is an extensive repository of research publications from conferences in the field of computational linguistics. We retrieved all publications between 2016 and 2024 with their bibliographic description, including the title, keywords, and abstracts. We have automatically filtered the publications based on a pre-compiled list of 18 search terms (e.g., 'MWE', 'multiword expression', 'phraseme', etc.). A list of 1,251 publications was retrieved, which were then checked by the authors.

3. *Europhras Conference Proceedings Repository*[7]. It provides lists of publications with relevant metadata. All publications after 2016 were checked. The resources retrieved overlapped with the identified resources from ELG and ACL repositories.

4. *Phraseology and Multiword Expressions book series*[8] (Language Science Press) established in 2017. The series includes books and collections addressing topics related to theoretical, computational, and empirical approaches to multiword expressions, including lexical resources. Several resources were identified in these publications that provided an extensive description of the linguistic information

[5] https://live.european-language-grid.eu/

[6] https://aclanthology.org/

[7] http://www.europhras.org/en/conferences

[8] https://langsci-press.org/catalog/series/pmwe

and representation of MWEs.
5. *Arxiv digital open access repository*[9]. It includes a wide range of scholarly articles in different areas. We searched in the 'Computer Science' category using the search terms list and identified several resources. While these mostly overlapped with previously identified resources, there were several new ones, particularly used in language processing applications.

In addition to the above, we asked community members working on MWEs for information on newly developed or updated resources not published in the examined repositories.

As noted, a systematic approach was adopted in this survey to identify and select resources related to MWEs. Inclusion criteria were defined to ensure that the reviewed resources fall within the scope of the survey and reflect the current state of MWE-related lexica that can be used in NLP tasks. The following inclusion criteria were applied: (i) date of creation, update, or publication of the resource, (ii) foreseen usage, (iii) type of lexicon (i.e., computational as opposed to lexica aimed at human users), and (iv) description of MWE entries. For comparison, only 45% of the monolingual and 66% of the multilingual resources in the PARSEME survey (Losnegaard et al. 2016: 2302–2303) are classified as MWE lexica; the most significant proportion of the resources are lists of MWEs. In the present survey, we exclude lists unless they are supplied with linguistic information such as lemma, syntactic description, semantic properties, etc.

Summing up, the selected resources contain MWEs as entries, focusing on syntactic, semantic, and other information relevant to their structure, meaning, and usage. Resources that are freely available or have academic licenses were prioritized to support collaborative and accessible research. Finally, the survey focuses on collections supporting NLP tasks involving MWEs.

3.2 The spreadsheet

The information about the selected resources was collected and organized into several sheets, each detailing various properties of the MWE resources, distributed as follows:

- The *General Information* datasheet captures high-level information about each lexicon, including its full and abbreviated names, publication date, language coverage, and availability. It also notes whether the resource is linked to a corpus and provides any additional comments or specific references for proper citation.

[9]https://arxiv.org/

- The *Linking to Corpus* sheet specifies whether the listed resources are linked to specific corpora, describing the types of corpora used, the source of the data, and the corpus size. It highlights resources where corpus integration plays a role in MWE representation.
- The *Lemma Representation* sheet provides information on whether and how lemmatization information is included within each resource, covering aspects such as morphological representation and syntactic information.
- The *Lemma Analysis* sheet provides a more detailed overview of how each lexicon represents lemmas (canonical forms) for MWEs, capturing various linguistic and syntactic features. It focuses on the nuances of lemma representation within each resource.
- The *Semantic Descriptions* sheet details the mechanisms employed for representing the meaning of MWEs, including glossing, compositionality ratings, and the encoding of semantic relations.
- The *Syntactic Descriptions* sheet outlines the syntactic properties of MWEs, indicating whether resources account for word order, inflection, or syntactic patterns.
- The *Resource Description* sheet provides an overview of each resource's size, maintenance, and licensing.

4 An overview of MWE lexical resources

The result of this survey is a list of 66 resources (compared to 107 reported in the PARSEME survey) dedicated to MWEs or containing MWEs, alongside other words. This list was reviewed and edited to ensure consistency, completeness, and precision. It records detailed information about each resource, including publication date, resource type, linguality (monolingual, bilingual, multilingual), acquisition method, and licensing information. These are extracted either from the paper documenting the resource or from the resource website, or observed via resource inspection. The results of the survey are openly available.[10]

This section provides an overview of the lexical resources included in this survey along the following axes: (a) time span, (b) intended or foreseen usage of the resource, (c) linguality type (i.e., monolingual, bilingual, or multilingual

[10]https://anonymous.4open.science/r/Lexical_Resources-BC18/

lexicon) and language(s) covered, (d) types of MWEs included, (e) acquisition method, (f) accessibility and type of license, (g) representativeness, as well as (h) linking of MWE lexica to other resources (corpora or other lexica).

4.1 Time span

The first inclusion criterion was the date of the creation, update, or publication of the resources, focusing predominantly on lexical resources produced after 2016, as our survey is a follow-up to the PARSEME one. Most identified resources are new; only three are enriched and updated. We also included several resources published before 2016 that did not feature in the PARSEME survey.

Figure 1 shows the number of resources reported in the PARSEME survey and our survey by year of publication. It can be seen that there was a peak in publishing resources in 2016, according to collective data from the PARSEME survey and ours. In the following years, a slower but steady trend is observed in developing new MWE resources.[11] The distribution of resources by year of publication is plotted against relevant EU-funded initiatives for reference: META-NET Project[12], PARSEME COST Action, Horizon 2020 ELEXIS Project[13], UniDive COST Action[14].

4.2 Intended usage

The main inclusion criterion was the intended or foreseen usage, as we were specifically interested in computational MWE lexica. However, we also identified lexical resources designed to serve both downstream NLP tasks and the needs and requirements of human users. The latter are less numerous than the former: from a total of 66 resources, 52 (78%) are computational, 13 (19%) are both computational and for human users, and only one resource is non-computational. The PARSEME survey results report the same distribution: most resources are for NLP usage, and only a few are for human use.

Additionally, we evaluated the usage of the resources. Fifty resources were broadly designated as applicable for NLP purposes, with compositionality rating being the most prevalent NLP task (eight resources). Six resources are meant for

[11] A possible explanation for the low numbers in 2021 is the limited number of conferences and forums for reporting research results due to the COVID-19 pandemic. The numbers for 2024 are expected to increase as publications from the second half of 2024 may not yet be included in the examined repositories.

[12] http://www.meta-net.eu/

[13] https://elex.is/

[14] https://unidive.lisn.upsaclay.fr/

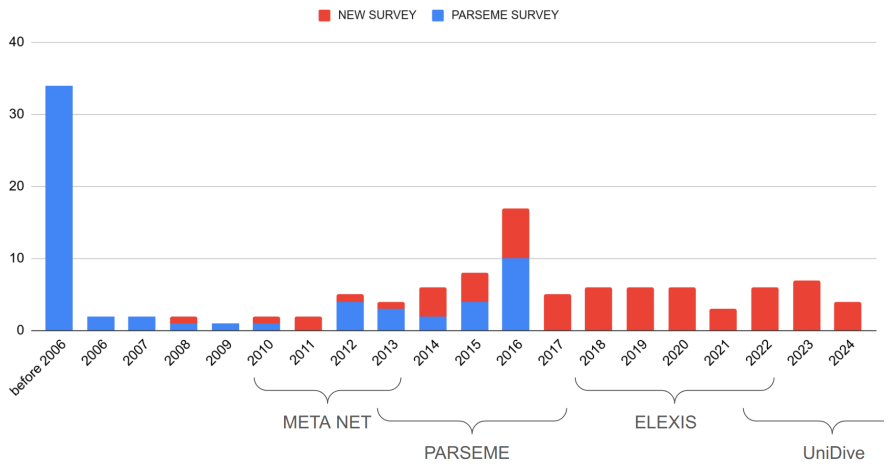

Figure 1: Distribution of resources by year of publication

human use. In the relevant documentation, the information about resource use was sometimes unclear (9) or absent (1).

4.3 Languages covered and linguality type

The linguality type of the resources refers to whether they are mono-, bi-, or multilingual. Of the selected 66 lexical resources, 51 (77.3%) are monolingual, 10 (15.2%) are bilingual, and 5 (7.5%) are multilingual. They cover 37 languages (42 including varieties) in total. For comparison, the PARSEME survey included 14 bi- or multilingual resources (13% of the total resource count). The multilingual resources in the PARSEME survey are predominantly multilingual lists of MWEs or translational equivalents compiled from lexical-semantic networks (such as WordNet or BabelNet) with scarce or no linguistic description, and, as mentioned before, such resources are not included in the current survey.

More precisely, we identified monolingual lexica for 24 languages. Below, we list these languages, indicating in brackets the number of lexica available when more than one: Arabic (AR) (2 lexica), Bulgarian (BG) (2 lexica), Croatian (HR) (2 lexica), Czech (CZ) (5 lexica), Dutch (NL) (3 lexica), English (EN) (5 lexica), Estonian (ET) (2 lexica), Finnish (FI) (1 lexicon), French (FR) (1 lexicon), German (DE) (1 lexicon), Modern Greek (EL) (3 lexica), Hebrew (HE) (1 lexicon), Irish (GA) (1 lexicon), Italian (IT) (1 lexicon), Lithuanian (LT) (1 lexicon), Polish (PL) (2 lexica), Portuguese (PT) (2 lexica), Russian (RU) (2 lexica), Serbian (SR) (2 lexica), Slovenian (SL) (3 lexica), Spanish (ES) (3 lexica), Swedish (SV) (1 lexicon), and

Yiddish (YI) (1 lexicon). Notably, two lexica feature MWEs specific to two varieties of Spanish spoken in Chile (ES-CL) and Argentina (ES-AR). A minority language, Pomak, is represented by one MWE lexicon.

Another ten lexica are bilingual, covering twelve languages (six of which do not appear in monolingual resources) and nine language pairs. Half of these are unidirectional from a source language to the target: English-French (EN-FR), English-Italian (EN-IT), English-Persian (EN-FA), Georgian-Modern Greek (KA-EL), Croatian-English (HR-EN), Polish-English (PL-EN); one resource is a bilingual dictionary that covers both directions, Basque-Spanish (EU-ES) and Spanish-Basque (ES-EU). One resource involves two languages, Bulgarian (BG) and Romanian (RO), linked using English (EN) as the pivot following the standard methodology for aligned wordnets. Additionally, one resource involves two language varieties that belong to the Indo-Aryan subfamily, namely Hindi (HI) and Marathi (MR) – yet, they are not aligned as translation dictionaries. Moreover, five resources are multilingual, covering ten languages in all (three of these languages appear neither in mono- nor in bilingual resources). The multilingual MWE resources vary only slightly in terms of the number of languages covered. One resource covers five languages, namely English (EN), German (DE), Italian (IT), Portuguese (PT), and Russian (RU), while another resource covers four languages: Japanese (JA), English (EN), Chinese (ZH) and Korean (KO). Two resources are trilingual; the first one includes English (EN), French (FR), and Portuguese (PT), and the second one includes English (EN), Chinese (ZH), and Japanese (JA). The final one includes one language as a source, Spanish (ES), with its varieties.

Our findings corroborate the observation by Losnegaard et al. (2016) that bilingual and multilingual MWE resources, including lexical ones, are rare. Despite years of research in this field, the scarcity of bilingual and multilingual MWE lexica remains a significant challenge. This limitation could impede research on MWE translation and cross-lingual NLP tasks.

4.4 Types of MWE lexica based on content

Both MWE-dedicated and MWE-aware lexica were identified. The former contains only MWEs of various types, such as verbal, nominal, or adverbial ones, as well as multiword named entities and terms. By contrast, general lexica that include MWEs alongside single-word entries are considered MWE-aware (or MWE-inclusive) lexica. They incorporate MWEs either as part of their macrostructure as independent entries or in their micro-structure as sub-entries under single-word main entries.

We also considered the type of MWEs in each resource, i.e., whether all kinds of MWEs are included or are limited to some specific type(s) (nominal, verbal, compound phrases, idioms, collocations, or some combination of these types). Terminological resources were excluded from this survey based on the assumptions that (a) terms are not consistently selected according to solid criteria for idiosyncrasy, and (b) no detailed MWE-related encoding is provided in pure terminological resources. However, we retained resources that either include terms alongside other types of MWEs (one resource) or handle multiword terms in a way that accounts for their idiosyncrasies.

We examined the MWE types that reference the morphosyntactic properties of the MWE and its function as part of speech (POS). While for 37.6% of the covered resources, all types of POS are declared to be included, for 29.4%, the POS is not specified. Of the remaining resources, those containing verbal MWEs are prevalent (17.7%),[15] and two are limited to verb-noun structures. The resources of nominal MWEs account for 7.1% of the total count, with no additional restrictions.

4.5 Acquisition method

The acquisition method is also noteworthy. Most resources were reportedly developed either semi-automatically (26 resources or 39.9%) or fully automatically (8 resources or 12.1%). By contrast, 21 resources (31.8% – approximately one-third of the lexical resources in this survey) were developed manually. All these three development methods are mentioned by Losnegaard et al. (2016), but without offering any quantitative data. To a certain degree, the distribution reported for our survey is to be expected, since many resources are compiled from pre-existing lexica or databases which were processed at least partially automatically. However, a large proportion of the resources (over 70%) required at least some level of manual work, which shows the difficulty of providing reliable and accurate linguistic descriptions of MWEs automatically.

It is important to note that information on the acquisition method was not readily available for some resources, indicating a need for further investigations on reported works on the resource.

4.6 Accessibility

The availability of resources is crucial as it directly impacts their accessibility and potential for their (widespread) use. Resources, free or free for specific purposes such as academic research, can foster greater collaboration and innova-

[15] This is probably correlated with the PARSEME focus on verbal MWEs only.

tion within the community. Based on the information from the developers, we identified 49 (74%) resources that fall into these categories. Other resources are available for a fee (three resources), which limits their accessibility. Additionally, for some resources, it remains unclear whether they are available at all (14 resources), further complicating their potential usage and integration into various downstream NLP tasks and applications. In the PARSEME survey, 95 resources (88% of the 107 resources) were found available, split almost evenly (46:49) between resources of unrestricted and restricted use, respectively.

Getting back to our survey, 21 resources (31.8%) are accessible through a dedicated link or platform, while 30 (45.5%) are available via specialized repositories, such as CLARIN, ELG, GitHub, LINDAT/CLARIAH.

4.7 Representativeness

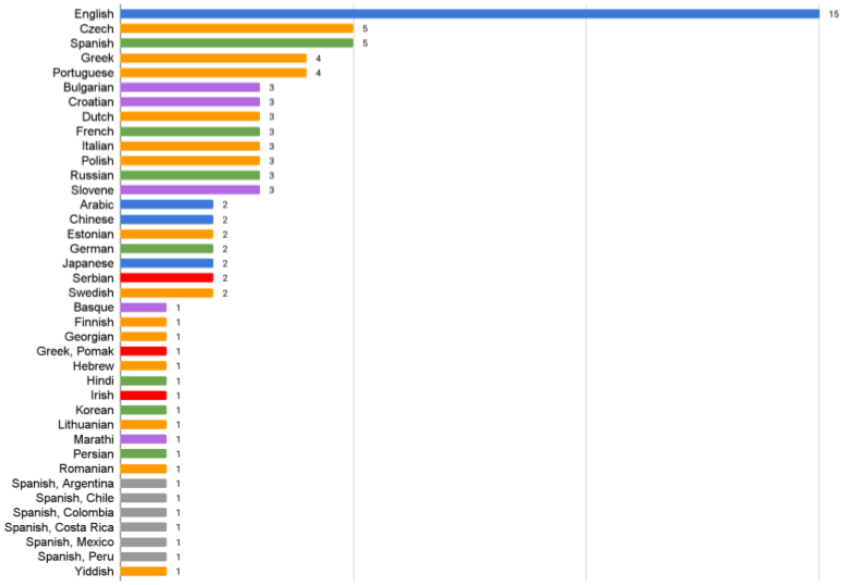

Figure 2: Distribution of different languages and the number of MWE resources they are involved in. The color shows the level of technical support: blue – Good; green – Moderate; orange – Fragmentary (higher); purple – Fragmentary (lower); red – Weak or none; gray – Unknown.

Not all languages are equally represented in the language resource landscape. In this regard, we attempted to examine whether the level of language representation correlates with the number of MWE lexica available for that language.

Hereon, we adopt the classification of languages presented by Maynard et al. (2022), the notion of Digital Language Equality (DLE) and the DLE metric defined by Gaspari et al. (2023). With respect to their overall state of technology support, we divide the languages into several categories: Good; Moderate; Fragmentary (higher); Fragmentary (lower); Weak or no support.[16]

According to Gaspari et al. (2023), DLE refers to the state where languages have the necessary technological support and situational context to thrive as living languages in the digital age. The DLE metric quantifies a language's digital readiness, its contribution to technology-enabled multilingualism, and its progress toward achieving DLE.

However, the ELE survey focused only on European languages and their level of representation in the digital world, leaving out languages outside Europe. To fill this gap and account for languages other than those spoken in Europe, we used a metric defined by Joshi et al. (2020), who use somewhat different considerations for their measurements. Namely, they consider world languages and their role in language technologies. They suggest that there exists a correlation between language typology and the level of language resourcefulness. In short, they divide languages into six classes arranged on a scale from 0 to 5 according to the available resources and data. We align the six-point scale of Joshi et al. (2020) to the five-point ELE scale: merging level 0, which implies a total lack of resources, with level 1, and aligning them to Weak or no support; level 2 – Fragmentary (lower); level 3 – Fragmentary (higher); level 4 – Moderate; level 5 – Good).[17] In this way, we obtained data on the level of support for some of the languages which are not in the ELE survey: Chinese-Mandarin (Good/5), Japanese (Good/5), Arabic (Good/5), Russian (Moderate/4), Korean (Moderate/4), Hindi (Moderate/4), Persian (Moderate/4), Ukrainian (Fragmentary (higher)/3), Georgian (Fragmentary (higher)/3), Hebrew (Fragmentary (higher)/3), Marathi

[16]Since the majority of European languages are of a Fragmentary level of support, according to ELE reports (https://european-language-equality.eu/deliverables/), we split the Fragmentary level into two levels – Fragmentary higher, which is closer to the Moderate support level, and Fragmentary lower, closer to the Weak or no support level. For example, Finnish is very close to Moderate level, Catalan would be on the very border between Fragmentary higher and Fragmentary lower, and all the other languages above this borderline would be Fragmentary higher (Polish, Swedish, Dutch, Portuguese, Italian, and Finnish).

[17]There are some discrepancies in the alignment for two languages, namely Irish (Weak or no support in ELE report and 2 in Joshi et al. (2020)) and Dutch (Fragmentary (higher) in ELE report and 4 in Joshi et al. (2020)). We decided to keep their ratings from the ELE report and acknowledge that the misalignment is small, only ±1 level. Also, Joshi et al. (2020) evaluates English at the same level as Spanish, German, and French, but we decided to keep the ratings from the ELE report. Full classification of languages by Joshi et al. (2020) is available here: https://microsoft.github.io/linguisticdiversity/.

(Fragmentary (lower)/2), Yiddish (Weak or none/1). Pomak is not classified, but there are very limited resources for it, and we assume its level of support is 'Weak or none.' Although Spanish is generally classified as 'Moderate,' we have no sufficient information about the level of support for its variants; thus, we classify them as 'Unknown.'

The overall distribution of languages in the reported resources with respect to their digital support is shown in Figure 2 (alternatively, the data are shown in Table 3 in the Appendix). Again, English is the best-represented language, appearing in nearly half of the language pairs as a source, target, or pivot language.

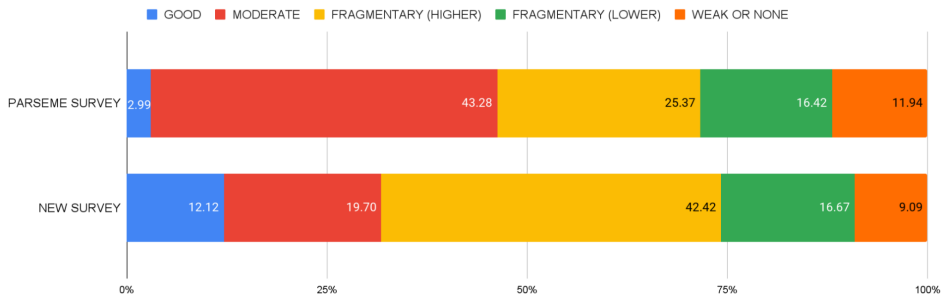

Figure 3: Distribution of resources according to level of technical support

Figure 3 shows the distribution of resources with respect to the level of technical support of the languages involved (for bi- or multilingual resources, we assign the level of representation for the language at the lower or the lowest level of support) in the PARSEME survey and the new survey.

The data clearly shows a general shift of focus since 2016 towards compiling MWE lexica for languages with fragmentary, weak, or no support, along with a continuing trend in developing MWE lexical resources for better-resourced languages. Furthermore, there has been extensive work on MWE lexica for non-European languages and language varieties, mainly varieties of Spanish.

Large corpora and rich embeddings remain scarce for low-resourced languages. This underscores the importance of reliable lexical data in facilitating the proper treatment of MWEs.

Moreover, the results show that resources for fragmentary lower-represented languages and fragmentary higher-represented languages alike are mostly linked to corpora or other data sources. By contrast, well- and moderately-represented languages tend to have lexical resources proportionally or equally linked to corpora and other data sources.

4.8 Linking of MWE lexica to other resources

Linking MWE lexica to corpora and other language resources would increase their applicability for various semantically oriented NLP tasks. Therefore, we further examined whether the identified lexica are linked to other language resources, such as corpora and other lexica (providing the name of the respective resource(s) where available). Of the lexical resources analyzed, only 22 (or 33.3%) are linked to a corpus, while the remaining 44 (66.7%) are not, as shown in Figure 4.

24.2% of the lexica covered in the survey are linked to other lexical resources (such as WordNet, BabelNet, or other computational dictionaries). As Figure 4 shows, a portion of the corpus-bound lexical resources is also linked to other lexical data sources.

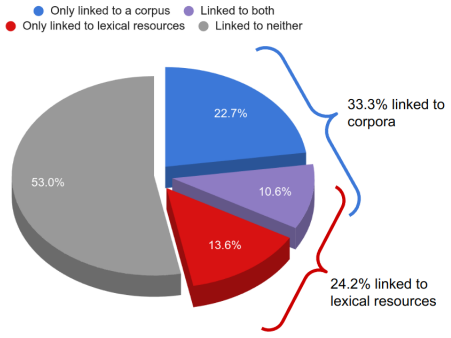

Figure 4: Distribution of resources according to their links to corpora and/or other lexical resources

Lexica linked to corpora are predominantly derived automatically (27.3%) or semi-automatically (50%), with only two cases (9.1%) of manually constructed lexica; in the remaining three cases, the method of compilation is unclear. No manually constructed MWE lexical resources are linked to other lexical data.

Overall, our analysis shows a deficiency in linking MWE resources to corpora and other lexical data. Corpora-linked MWE resources are predominantly automatically derived MWE lists with little or no linguistic description and no confirmed MWE status (as stated previously, such lists are not included in the present survey).

In this section, we have described in detail the 'macroscopic' properties of MWE resources. We will discuss the findings of this section in combination with more information about the 'internal structure' of the resources in Section 8. In the remainder of this text, we present the situation regarding the forms that

are used to represent MWEs and how meaning is encoded in the resources, and reflect on the recent technological landscape as it is formed by LLMs.

5 The MWE identifier

In this section, we consider a MWE entry as an entity defined by a 'meaning' and a non-empty set of MWE forms (Skoumalová et al. 2024, Markantonatou et al. 2024), where a MWE form is a (partially) ordered set of independent words, or a multitoken word. The term 'word' is used here as a cover term for lemmas and inflected types. In each MWE form, these words stand in a fixed set of dependence relations, including those that characterise multitoken structures in a particular language.

The need for a representative form of the MWE arises because many MWEs inflect and are subject to variation phenomena. Typically, one MWE form is used to represent and identify a MWE entry. We will call this MWE form the *MWE Identifier (MWEI)*.[18] However, resources that take into account variation may end up with a disjunction of MWEIs (see Section 5.4).

A MWE entry may also include a bundle of lexical, morphological, syntactic, semantic, pragmatic, and distributional pieces of information about a MWE that is not represented in its MWEI.

5.1 Types of MWEI

There are two types of MWEI. The *Lemma MWEI* applies the traditional notion of lemma on MWEs. A lemma is one form of a word that is conventionally considered to systematically represent the set of possible forms of the respective word. In the remainder of this section, we will discuss issues related to the formulation of this representative form.

The *Abstract MWEI* is a representation of the MWE that is not necessarily identical to one of its usable forms. In this approach, all the components of the MWEI are lemmas. Sometimes morphological information is attached to the lemmas in the Abstract MWEI and some other times a list of the possible forms of the MWE is given. The Abstract MWEI, although consisting of lemmas, may not be a usable form of the MWE because:

- Very often, some of the components of the MWE may remain morphologically fixed across the instances of the MWE, and others may inflect.

[18] Other terms for this type of identifier that can be encountered in the literature on this topic are 'canonical form' and 'MWE lemma'.

- The inflected components may be able to assume all or only some of the grammatically permissible options in their inflection paradigm.

Resources employing a Lemma MWEI may also provide its lemmatized form. While the lemmatized form of the Lemma MWEI is identical in structure to the Abstract MWEI, it is important to note that it does not function as the representative instance of the MWE.

In our collection of resources, 52 resources used Lemma MWEIs and 6 Abstract MWEIs, while it was not clear what 8 resources have used as a representation form.

Below, we discuss in extent how resources handle the Lemma MWEI.

5.2 Definition of lemma MWEI

Defining a Lemma MWEI is not a trivial task. Such a definition mustaccount for the inflectional relations among the various usable forms of a MWE and a set of phenomena that we collectively call 'variation'. In this section, we discuss the issue of inflectional relations. Some general tendencies are observed in the studied resources; however, there is a lot of room left for approaches shaped up by the particularities of each language.

Typically, a Lemma MWEI contains the morphologically fixed components of the MWE as they are and uses lemmas for the non-fixed ones. For instance, all instances of the (EN) expression 'to kick the bucket' contain the morphologically fixed sequence 'the bucket' while the verb 'to kick' may assume any type in its inflectional paradigm. Moreover, many function MWEs (adpositions, conjunctions, determiners) and adverbial MWEs have one MWE form where none of the components can be inflected, and any word order permutation cancels the idiomatic meaning. This unique form is used as the Lemma MWEI. Two examples of unique forms are given below:

1. The fixed expression (LA) *mutatis mutandis* and its equivalents (EN) *all being considered* and (EL) τηρουμένων των αναλογιών, Lit. respected.PART.FEM.GEN.PL the.DET.FEM.GEN.PL analogy.NOUN.FEM.GEN.PL; while in all languages the translation equivalent expressions function as adverbs, their fixed structures are very different.

2. Antunes & Mendes (n.d.) do not reduce expressions to a lemma form; instead, they preserve the surface morphological form of the verb, especially in cases where the expression conventionally appears in a fixed form, such as cooking recipes, e.g., (PT) *leve ao forno* 'put in the oven'.

Some languages use infinitives as verb lemmas. However, since a MWE entry is a coupling of meaning and form, it is often the case that the subject of verbal

MWEs is represented by a suitable pronoun that also functions as a carrier of semantic features such as '+human' and '-animate'. However, structures where a pronoun is attached to an infinitive as its subject are often grammatically unacceptable. An approach adopted in DUCAME (NL) (Odijk & Kroon n.d.) is to use a specific finite form of the verb in the Lemma MWEI rather than the infinitive.

Components that inflect but assume only some of the available choices of their inflectional paradigm also occur (some choices may be excluded by the syntactic structure of the MWE form). Different strategies are reported for these cases. For instance, the verb head of the (EL) expression μας χαιρέτησε, Lit. usPRON.ACC.PL wave.3SING.IND.PAST.PERF, 'died' appears only in past tenses and in the third person. In such cases, one may use in the Lemma MWEI the form that is conventionally closest to the lemma. For instance, IDION (EL) (Markantonatou et al. 2019) uses the verb form χαιρέτησε3SING.IND.PAST.PERF.

Since the Lemma MWEI may contain the lemmas of words, issues related to the definition of a lemma are important. For instance, Antunes & Mendes (n.d.) uses different lemmas for a verb and its past participles. Also, at the moment, it is still debatable if forms of nouns, adjectives, adverbs and verbs expressing some type of degree modification should be assigned the lemma of only the base form or that of the base form plus the affix (Odijk & Kroon n.d.).

In sum, as regards the form of the lexicalised parts of a MWE, there is a general consensus in the definition of the Lemma MWEI among resources. Problems occur when a language uses the infinitive form as a verbal lemma and when a MWE has a gappy inflectional paradigm. Additionally, there are disagreements regarding the lemma form of diminutives, adjectival degrees, deverbal nouns and participles.

5.3 Free complements and the Lemma MWEI

MWEs may support one or more free phrases that complement the meaning of the MWE by instantiating core syntactic functions (henceforth the 'free arguments' of the MWE), as well as possessive and oblique complements. There seems to be some convergence regarding the representation of the free nominal complements. There is no convergence regarding the representation of other types of free complements, such as clausal ones.

Resources may encode free nominal arguments and free possessives as part of a Lemma MWEI. Generally, indefinite pronouns are used as variables in the possible nominal complements of the MWE. The pronouns encode the morphological requirements imposed on the free arguments of the MWE, such as case and number. Also, the semantic distinctions they can encode in some languages are significant, such as the distinction between human vs nonhuman entities and

animate vs inanimate entities. It is important that pronouns serve as indicators of binding phenomena, which are sometimes necessary to obtain the idiomatic meaning. In general, the usage of indefinite pronouns as variables that range over the free complements of a MWE gives a more precise picture of the semantic and syntactic structure of the MWE and makes the resource self-standing (Odijk & Kroon n.d.).

Several resources do not encode information about the free arguments on the Lemma MWEI. Indeed, there are reasons for this choice. The first reason is that MWEs are identified mostly from their non-variable components (although the semantics of the free complements plays a significant role in the identification of a MWE). A second reason is that languages with infinitives may have to choose other representative forms than the conventional verb lemma in order to encode free arguments with a verbal Lemma MWEI; the problem was discussed in Section 5.2. Finally, if MWE resources can benefit from linking with a general lexicon of the language, it may be assumed that some information on free arguments can be inherited from the head of the MWE, e.g., the verb head of a verbal MWE. For instance, the Lemma MWEI may be (EN) *to give a hand* because the free argument *to someone* is assumed to be prior knowledge about the valency properties of the verb *to give*. This approach cannot be applied to languages that do not have large general lexica. Furthermore, given the idiomatic meaning of MWEs, linking with the entries of nonidiomatic language may present challenges; therefore, relevant provisions should be foreseen in the linking mechanisms.

In sum, there is a consensus about using indefinite pronouns to represent the free nominal complements of MWEs in the Lemma MWEI. There is no consensus about representing free clausal complements of MWEs in the Lemma MWEI. There are open issues regarding the number of free arguments that should be included in the Lemma MWEI.

5.4 Variation and the lemma MWEI

MWE forms are subject to variation. Which phenomena are labeled as 'variation' is still debatable (Skoumalová et al. 2024). Most resources do not encode any variation.

Variation is interesting in the framework of a discussion about the Lemma MWEI because we have defined a MWE entry as the coupling of one meaning with a non-empty set of forms, and variation is the reason why this set may contain more than one form. By definition, only forms that can be coupled with the same meaning can be considered in this discussion, but even this constraint leaves room for different approaches. Before progressing further, it is useful to

clarify that the selection of variation phenomena discussed here only serves the purposes of this survey and does not aim to be a definition of 'variation'.

Since there is no strict definition of variation or any agreement on the treatment of specific phenomena, each resource adopts its solutions for each type of variation. The types of variation discussed here are lexical variation, spelling variation, optionality of a component of the Lemma MWEI, and inflectional variation of the fixed components. Syntactic variation is discussed in Section 5.5.

5.4.1 Lexical variation

Lexical variation occurs when a fixed component of the MWE can be replaced by a synonym, a hypernym or a hyponym without a change of meaning. Resources may:

- group lexical variants in a set of disjunctive Lemma MWEIs (Grégoire 2010, Antunes & Mendes n.d., Skoumalová et al. 2024) or
- define a separate entry for each lexical variant and semantic relation among the entries, such as synonymy, among the entries (Markantonatou et al. 2019, Leseva et al. 2024, Giouli et al. 2024).

The problem with the disjunctive Lemma MWEIs is that each variant may have its own syntactic, stylistic and/or distributional particularities that must be encoded separately. LEMUR (Skoumalová et al. 2024) is a resource that allows encoding such particularities with a disjunctive Lemma MWEI approach. On the other hand, the problem with the separate entry approach is that the user cannot directly see the affinity between the variant Lemma MWEIs without resorting to some additional feature of the resource, for instance, a facility that enables the user to see the synonyms of a MWE (IDION, Markantonatou et al. 2019) or the fact that the expressions share the same synset ID, for the resources aligned to wordnets (e.g., the Bulgarian-Romanian dictionary, Leseva et al. 2024), or the same frame ID, in case of the resources aligned to framenets (e.g., the Greek FrameNet including also MWEs, Giouli et al. 2024).

Another type of variation of the lexical type that might be considered occurs when one or more functional words, which are fixed components of the MWE, are substituted for other functional words or adverbials, or are omitted (determiners are often omitted) with no change of meaning. In such cases, most resources adopt the disjunctive Lemma MWEI approach, since there seem to be no distributional or other particularities characterising the variants.

5.4.2 Other types of variation

Spelling variation: This occurs when spelling variants of words are used and the resource encodes them. Such is the case of Modern Greek (EL); several (EL) words have spelling variants defining disjunctive Lemma MWEIs.

Inflectional variation of the fixed components: The morphologically fixed components of the Lemma MWEI may inflect with no change of meaning, e.g., θολώνουν τα πράγματα, Lit. blur.3RD.PL the.PL.NOM thing.PL.NOM, θολώνει το πράγμα, Lit. blur.RD.SG the.SG.NOM thing.SG.NOM, 'a situation becomes unclear or uncertain'. Resources encoding this type of variation may define disjunctive Lemma MWEIs or use the lemma of the variant word (Antunes & Mendes n.d.).

Derivation versus inflection: The same meaning can be related with constructs that, depending on the theoretical approach adopted by the developers of the resource, belong to the derivation or the inflection paradigm of the head of the expression. For instance, a verbal MWE and its nominalised forms, or the diminutives of nouns and adjectives can be considered to be represented by the verb and the noun or adjective lemma, respectively, or to be represented by their own lemmas. Provided that the original idiomatic meaning is not affected, resources may group such constructs as variants under one entry; for example, IDION (Markantonatou et al. 2024) groups diminutives under the basic noun or adjective lemma. Alternatively, they may define different entries; for example, DUCAME (Odijk & Kroon n.d.) defines different entries for MWEs, including diminutives, because it assigns a new Lemma MWEI featuring the diminutive.

Still, in some languages, such as the Kartvelian ones, the Slavic languages, and Modern Greek, the use of diminutives sometimes implies a polarity or a degree modification of the MWE meaning. In practice, the treatment of degree modification varies across resources, depending on whether it is regarded as a shift in the idiomatic meaning (and thus warrants a new entry) or is encoded within the non-modified MWE entry.

In summary, lexical variation occurs when lexicalized words are substituted for other words or omitted without affecting the idiomatic MWE meaning. There is no widely accepted definition of 'variation', and resources usually do not encode it except for the variation of determiners. Encoding methods also vary widely.

5.5 Syntactic aspects and the Lemma MWEI

Lemma MWEIs are generally regarded as grammatical structures in their languages, and as such, they are subject to the respective grammar rules. However,

sometimes MWEs may block some rules that could apply to them. Conversely, they may allow for phenomena that normally should not be encountered with these structures.

Below, we discuss the meaning-preserving word-order permutations and the modification of the fixed components of a MWE (bearing in mind that modification changes the meaning of the MWE to which it applies). These are the most documented syntactic operations on Lemma MWEIs.

Word order. The Lemma MWEI normally encodes the less marked available word order in a language. Word-order permutations that are the result of the grammar rules of a language are not encoded in most of the resources. Exceptional, meaning-preserving word orders are encoded as variants of the MWEI (Skoumalová et al. 2024). For instance, the nominal MWE form (SR) *Farenhajtov stepen* 'Fahrenheit degree' consists of an adjective followed by a noun, which is the natural order of an adjective and a noun in a Serbian MWE. However, the reverse order *stepen Farenhajtov* is also allowed for this MWE, as for many other (but not all) adjective-noun combinations in Serbian. Therefore, such information is included in the MWE description.

Conversely, some MWE forms may not allow for otherwise grammatical word permutations. For instance, in Modern Greek, the normal order is verb-object, however the MWE (EL) αγρόν ηγόραζε, Lit. field buy.3SING.PAST.IND 'one does not care' appears only in the order object-verb. Given this complicated picture of word permutations in MWEs, certain resources encode all the attested meaning-preserving word order permutations following different strategies. LEX-MWE-PT (PT) encodes them as different entries (Antunes & Mendes n.d.). IDION (EL) encodes word order possibilities with usage examples retrieved from Google searches (Markantonatou et al. 2019).

Modification of the lexicalised words in the Lemma MWEI. In several MWEs, a word or a phrase may modify its lexicalised components. When the modifier applies to the whole idiom as a unit, we talk about external modification, while internal modification occurs where the modifier applies only to the word on which it depends. Example 1 shows the former case in Romanian, i.e., the adverb *ușor* modifies the whole MWE, while example 2, the latter, i.e., the adjective *dure* modifies only the noun *cuvintele* in the MWE.

(1) Romanian
Copilul nu și -a **retras** **cuvintele** ușor.
Child-DEF not self's-REFL.DAT.3SG -has withdrawn words-DEF easily.
'The child did not withdraw his words easily.'

(2) Romanian
Copilul nu și -a **retras** **cuvintele** dure.
Child-DEF not self's-REFL.DAT.3SG -has withdrawn words-DEF harsh.
'The child did not withdraw his harsh words.'

Several resources indicate the positions within a Lemma MWEI where a modifier can be inserted, and probably list some such modifiers (which are sometimes subject to strong selection). Other lexica adopt a different approach (Skoumalová et al. 2024, Leseva et al. 2024): when modification follows the syntactic rules of the language, no information is provided about it in the lexicon; only cases when such regular modifications are blocked are recorded.

5.6 Summing up about the Lemma MWEI

The Lemma MWEI is the most popular choice in the studied resources. There is some convergence on the definition of the Lemma MWEI, but open issues still exist regarding lemmatization strategies and the representation of free MWE complements. Lexical variation also motivates a variety of approaches, which have an impact on the definition of the Lemma MWEI. Overall, there is still room for discussion about further conversion among resources regarding the definition of the Lemma MWEI.

6 Meaning representation

For the correct interpretation and use of MWEs, it is of the utmost importance to describe them, given their usually non-compositional meaning. Though not the focus of our work here, we remind the reader about the numerous dictionaries of MWEs that have been published (on paper or electronically) only to explain their meaning, either through a paraphrase, definition, synonym, or by describing the historical circumstances of their coinage. Such dictionaries are all catering to human usage, thus implying that the meaning of MWEs is the only challenging aspect for a (native) speaker.

Definitions are usually provided in the language of the MWE. In addition, some resources provide translations into a widely spoken language, usually English (Markantonatou et al. 2019, 2024, Blagus Bartolec et al. 2024) and sometimes French, too (Markantonatou et al. 2019, 2024). Translations may be MWEs in

the target language or, more often, descriptions of the meaning of the MWE in the target language. In the case of aligned resources (see Section 6.1), translation equivalents can be offered in aligned languages.

Besides meaning description in the form of a definition, dictionaries may also represent the use of MWEs by providing one or more examples, either extracted from authentic texts or created by the lexicographer.

6.1 Types of meaning representation

With respect to the lexical resources included in this study, the representation of the meaning of the MWE is considered relevant. In addition to clarifying the meaning for the human user, the definition is meant to distinguish between different possible senses of the same MWE, to keep track of MWEs with the same meaning (i.e., synonymous ones), and, more recently, to improve LLM-based translation.

The resources we studied have used the following ways to represent the meaning of MWEs:

1. *Informal gloss*: this is a phrasal description of the meaning (Skoumalová et al. 2024). Informal glosses are not expected to satisfy the restrictions of analytical definitions, as they are not (necessarily) formulated by lexicographers.[19]

2. *Paraphrase:* paraphrasing has been proposed as a means for the semantic representation of idiomatic expressions with non-compositional meaning during lexical encoding (Villavicencio et al. 2004). Paraphrases of MWEs have been claimed to produce better representations for sentences that contain idiomatic expressions (Wada et al. 2023). In their seminal work, Bhagat & Hovy (2013) define paraphrases as sentences or phrases that convey the same meaning using different wording. According to Beaugrande & Dressler (1981: 50), paraphrases offer "approximate conceptual equivalence between outwardly different material". In this context, a linguistic element P – whether word, phrase, or sentence – is a paraphrase

[19]In lexicography, an analytical definition is an intensional definition. Intensional definitions contain the necessary and sufficient conditions for the usage of the defined word. They contain the 'genus proximum' (i.e., the more general word than the defined one) and 'differentia specifica' (i.e., the characteristics that make the defined word unique among other possible words sharing the same genus proximum). Another type of intensional definition is the formulaic definition, in which the representation of meaning is made in a way that describes the syntactic behavior of the defined word.

of another element *A* if *P* and *A* can be mapped onto the same meaning *M*. In a sense, the semantic relation between two or more paraphrases at the phrase and sentence level is equal to synonymy at the lexical level. Mel'čuk (2012) considers paraphrases to be linked via a synonymy relation and they can, therefore, be mapped onto the same sense or *SemS*. Like synonymy, paraphrasing is not absolute, and we usually speak about quasi-paraphrases. However, paraphrasing is not a completely unstructured phenomenon. From a purely lexical perspective, (quasi-)paraphrases are obtained using specific devices: synonym/antonym substitution, actor/action substitution, general/specific substitution, etc. (Bhagat & Hovy 2013). In this regard, a paraphrase is distinct from a lexicographic definition, hence from an informal gloss.

3. *Glosses taken over from other resources* such as WordNet, FrameNet, etc.
    - WordNet or wordnets[20] – synsets[21] are always assigned a gloss and sometimes even a usage example; the glosses in the Bulgarian-Romanian aligned MWE lexicon (Leseva et al. 2024) are taken from the wordnets of these languages (Bulgarian and Romanian, respectively).
    - FrameNet (Baker et al. n.d.) or framenets[22] – each frame contained therein has a definition applicable to all lexical units pertaining to the respective frame: Giouli et al. (2020, 2024) present a FrameNet for Modern Greek which also contains MWEs, treated just like any lexical unit in such a resource, thus benefiting from a rich semantic description of their arguments along with the semantic roles they assume.
    - BabelNet (Navigli & Ponzetto 2012) is a large multilingual semantic network integrating rich information – lexical, semantic, and encyclopedic – from WordNet and Wikipedia. It relies on WordNet's structure organised around concepts (a concept corresponds to a synset – a set of synonyms) and the semantic relations linking them. Further,

[20] We use the form *WordNet* to refer to Princeton WordNet (Miller 1995, Fellbaum 1998). The common noun *wordnet* is used among specialists to refer to any language resource that follows the organization principles of Princeton WordNet.

[21] The nodes of the lexical network, i.e. of the wordnet, contain synonym sets (called synsets), which are the possible lexicalizations of a concept in the respective language.

[22] Just like in the case of the pair of words *WordNet – wordnet*, from the name of the resource developed by Baker et al. (n.d.) at Berkeley, i.e. *FrameNet*, the common noun *framenet* is used to refer to any resource that follows the same organization principles.

lexicalisation of concepts in different languages is provided, either harvested from Wikipedia or obtained through machine translation. BabelNet's main purpose is to enable knowledge-rich, graph-based Word Sense Disambiguation both in monolingual and multilingual settings. MWEs are included alongside single lexical units in each synset. The cross-lingual links allow to extract a multilingual translational dictionary of MWEs. Besides POS, no other linguistic description of MWEs is provided. CollFrEn (Fisas et al. 2020) is a lexicon of collocations in which each component is assigned a BabelNet sense.

When the definition is assigned from a wordnet, the MWE is included in a network of semantic relations. Since in wordnets definitions are associated with synsets, the MWE inherits the semantic relations of the synset to which it belongs. Figure 5 (Leseva et al. 2024) shows that a synset[23] (which may contain MWEs) enters several semantic relations rendered as arrows in the figure. Thus, MWEs enter semantic relations with other MWEs or with words in the language. What is more, whenever the respective wordnet is aligned to Princeton WordNet, this puts the MWE in a multilingual context, offering it semantically related words or MWEs in those languages (cross-lingual synonyms, etc.) and the possibility of being used in multilingual applications. For example, Figure 6 (Leseva et al. 2024) shows the aligned synsets in Princeton WordNet, the Bulgarian Wordnet and the Romanian wordnet, all containing MWEs lexicalizing the same meaning.

Within a framenet, MWEs are assigned to semantic frames, i.e., conceptual structures to which MWEs belong as lexical units and which specify the corresponding frame elements (e.g., Agent, Experiencer) realized by their arguments of the MWEs. Given that various relations are defined with other frames (Inheritance, Perspective-on, Using, Subframe, Precedes, etc.), MWEs, just like any lexical unit, participate in these relations.

4. *Lexical functions.* In his dictionary, which is general, not a MWE-dedicated one, unlike all the other lexica we discuss here, Mel'čuk (2006) treats MWEs of the type 'idioms' and 'quasi-idioms' as lexical entries, which is an uncommon practice in lexicography. Their meaning is described, alongside a formalized definition, using various lexical functions, encoding their synonyms, hypernyms, derivatives, collocations, etc. This way of defining

[23]The rectangles contain the corresponding synsets in three languages: English, Bulgarian and Romanian; this is possible given that they are aligned at the synset level and the Bulgarian and the Romanian wordnets borrow the structure from Princeton WordNet.

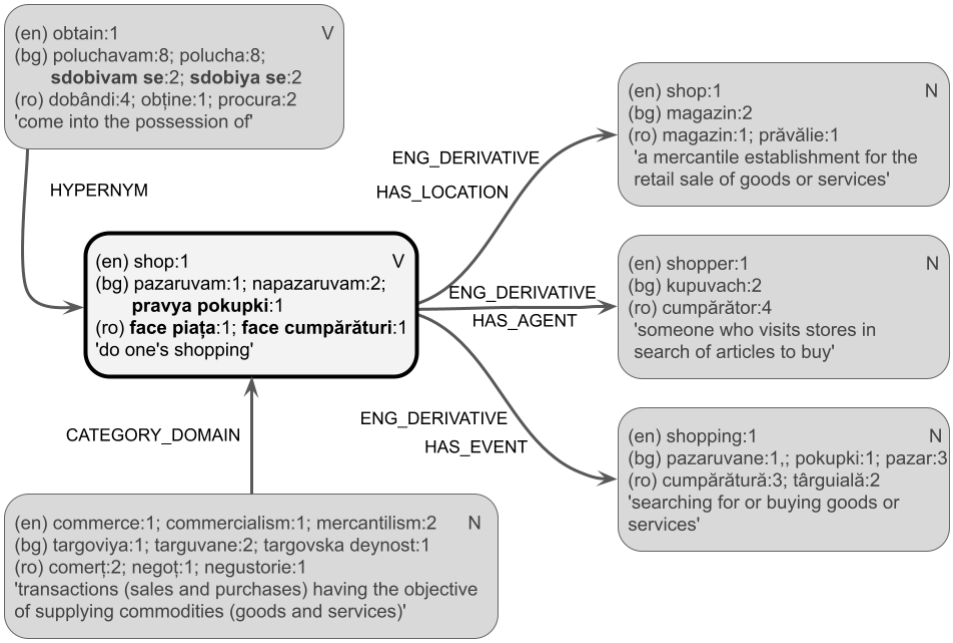

Figure 5: Semantic relations in wordnets, involving also MWEs (Leseva et al. 2024).

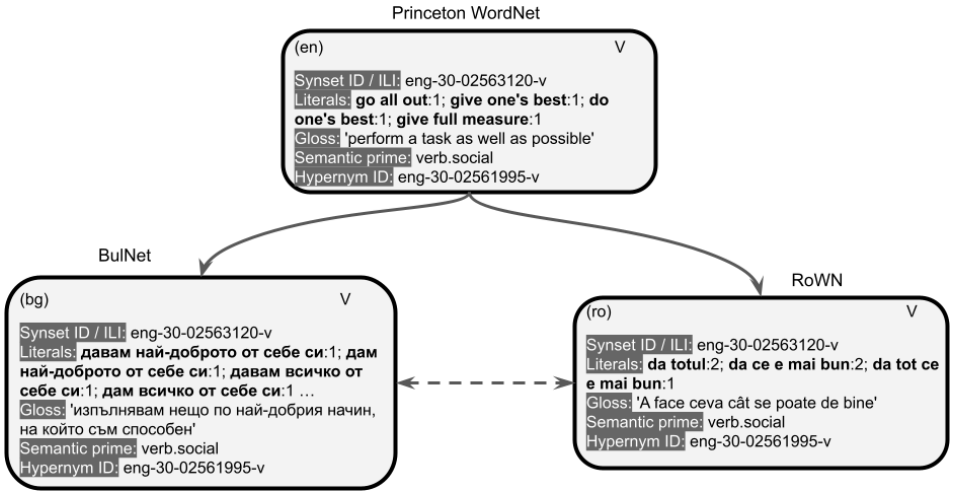

Figure 6: Corresponding synsets containing MWEs in interlinked wordnets (Leseva et al. 2024).

the meaning of MWEs of the collocation type is found in CollFrEn (Fisas et al. 2020): syntagmatic and paradigmatic lexical functions are used to describe the meaning of collocations.

5. *Vector space representations.* A more recent way of representing the meaning of a MWE, though in a format readable only by machines, even if interpretable by human specialists as well, is the vector space one. The typically non-compositional meaning of MWEs implies the non-compositionality of the vectors for each individual component of the MWE in representing their meaning in the vector space. This was proven even for collocations: vectors dedicated to the whole MWE provide the appropriate meaning representation, not the composition of the vectors of the individual words in the MWE (Fisas et al. 2020). In this case, vector space models have treated MWEs as individual tokens to generate static embeddings. These embeddings capture both the complex nature of MWEs and their potential semantic overlap with similar expressions. To encode the non-compositional meaning of an expression for MWE identification purposes (Liebeskind & HaCohen-Kerner n.d.), the meaning of the candidate's constituents can be represented by vectors in the same semantic space. Due to the idiomatic nature of MWEs, we expect the similarity between the vectors of words in a non-MWE to be greater than that of the words in a MWE. Thus, vector representations can help distinguish MWEs from non-MWEs by capturing the degree of compositionality in their constituent relationships within the semantic space.

6. Certain resources assign *compositionality ratings* to MWEs. These resources are used as benchmarks for systems that predict whether a set of words can be assigned a compositional meaning or a non-compositional one, in which case a MWE is identified. Reflecting a rather psycholinguistic perspective, the lexicon of English and Italian idioms (Pagliai 2023) contains annotations of two types of what is called *variables*: experience-based ones and content-based ones. Experience-based variables include the familiarity of an idiom (i.e., how frequent it is in a speaker's language experience), its meaningfulness (i.e., "the perceived, subjective knowledge of the figurative meaning of an idiom") and its objective knowledge (i.e., "the actual knowledge of the figurative meaning of an idiom", Pagliai 2023: 4). Content-based variables include literal plausibility (i.e., "the possibility that the literal-compositional meaning of an idiom denotes an event consistent with our world knowledge"), decomposability (i.e., "the possibility

of decomposing the figurative meaning of an idiom and associating the parts with its literal constituents") and transparency (i.e., "the possibility of tracing a synchronic relationship between the literal and figurative meaning of an idiom", Pagliai 2023: 2–4).

In addition to the description of the MWE meaning with some type of text, some resources document various semantic properties:
a. *Lexical semantic relations among MWEs*, such as synonymy, antonymy and n-tuples of MWEs that are related via causative/non-causative and aspect variation are also encoded by a small number of resources. Blagus Bartolec et al. (2024) group Croatian collocations in thematic fields.
b. The Japanese verb lexicon (Kanzaki & Isahara 2019) encodes information about *semantic roles* selected by the documented MWE.

Table 1 shows the distribution of resources according to the meaning representation technique they adopt. We notice that many resources (45%) do not encode the meaning of MWEs in any way, while for those in which the meaning is described, the preferred way is by informal glosses (24% of all resources and almost half of those that contain the MWEs' meaning description).

Table 1: The number of resources adopting the types of meaning representation techniques. (The lexicon which contains vector space representations for the MWEs also uses lexical functions to represent MWEs' meaning, thus the sum of the resources exceeds 66.)

Types of meaning representation	# of resources	% of resources
informal gloss	16	24.2
paraphrase	5	7.6
glosses from other resources	8	12.1
lexical functions	3	4.5
vector space representation	1	1.5
compositionality ratings	4	6.1
no meaning representation	30	45.5

In conclusion, the meaning of a MWE is primarily captured via text-based descriptions either inherited from general lexica or newly-written as an informal definition or a paraphrase. The MWE may be integrated in a network of semantic relations such as inheritance relations, lexical semantic relations such as equivalence (synonymy) and antonymy relations, as well as a set of more specific semantic properties such as semantic roles selected by the predicates and MWEs

related through causative / inchoative and aspectual variation. Certain resources assign compositionality ratings to MWEs and serve as benchmarks for systems that identify non-compositional structures in corpora. Finally, only one resource assigns embeddings to MWEs.

6.2 Examples of MWE use in dictionaries

Dictionaries may exemplify the meaning of an entry by providing one or more usage examples, usually extracted from a corpus, the Web or created by the lexicographer. When the dictionary is aligned with a corpus, the MWE entries are associated with their occurrences in the respective corpus. A limitation of corpora is that they must be sufficiently large and representative of diverse language uses to capture occurrences of MWEs. This requirement is further amplified when flexible or non-canonical usages of MWEs are also considered. In a sense, the data problem of linking lexica with corpora is similar to that of LLMs (see also Section 7), in particular because large resources usually include texts that contain few or no cases of colloquial speech, where the use of MWEs is usually more creative.

Usage examples become increasingly interesting because state-of-the-art NLP techniques rely on corpora. The PARSEME shared tasks (Savary et al. 2017, Ramisch et al. 2020) have relied on usage examples. Zeng & Bhat (2022) and Ide et al. (2024) use datasets of MWE usages to fine-tune a model for better MWE identification results. Recent attempts to translate verbal MWEs with LLMs showed that prompting a small LLM with pairs of usage examples in the source language and their translation in the target language improves translation results significantly (Dimakis et al. 2024).

In the set of resources examined here, 37 out of 66 provided one or more usage examples. Most of them (32, i.e., 48%) extract examples from corpora; of the rest, one resource uses recordings, three include constructed examples, and one lexicon draws its examples from FrameNet.

6.3 Summing up on the meaning representation

Although the most noticeable characteristic of MWEs is their lack of semantic compositionality, the MWE lexica developers may well tend to leave out their meaning description. One possible reason for this is the human-centred nature of this information when provided in the lexicographic style, where "lexicographic" is to be understood in a wide sense (traditional lexicography, wordnet-like, etc.). As the NLP-oriented lexica are designed to serve their contemporary technology,

the encoding of MWE forms was a more salient goal for symbolic and statistical NLP than meaning descriptions. State-of-the-art NLP techniques make heavy use of embeddings; in this context, vector space representations could prove their effectiveness. However, these are still only sparsely represented in the resources, given the important technical difficulties associated with them, e.g., lack of adequate resources from which they can be extracted and lack of storage space for such large amounts of data.

On the other hand, as LLMs have a dominant role in the NLP stage, textual representations of the MWE meaning such as usage examples (Dimakis et al. 2024), informal glosses, paraphrases or glosses taken over from long-standing resources (such as WordNet, FrameNet, BabelNet), may again present themselves as potential aids for downstream tasks such as machine translation with LLMs (Li et al. 2024). Exactly this point leads the discussion to Section 7 about the usefulness of lexica in the LLM era.

7 Are MWE lexica necessary with LLMs?

Given the recent important developments in AI and NLP in particular, the interaction between LLMs and lexical resources for MWEs is an issue. Despite the profound ability of LLMs to represent various types of linguistic knowledge, questions remain about how effectively these models handle MWEs. This section aims to provide a first picture of the efforts to better understand how to handle MWEs with LLMs; some of these efforts have used lexical resources.

Miletić & Schulte im Walde (2024) survey existing research in three key areas: (i) whether transformers can adequately capture and generalize the meanings of MWEs, (ii) which layers, tokens, and contextual elements are most influential in representing MWEs, and (iii) how the linguistic properties of MWEs affect the quality of these representations. The authors suggest that while transformer models can represent the meanings of MWEs to some extent, their ability to generalize across different types of MWEs remains inconsistent, and there is a need for further optimization and task-specific fine-tuning to improve these representations. The position of Miletić & Schulte im Walde (2024) on the limitations of transformer models in generalizing MWE meanings is shared by other researchers, who demonstrate that processing idiomatic or non-compositional expressions remains a challenging task for these models (Liu et al. 2024, Phelps et al. 2024, Yu & Ettinger 2020), although more sizeable LLMs or LLMs endowed with targeted (attention) mechanisms may return better results.

LLMs trained on large corpora have demonstrated their ability to capture the meaning of words and phrases through contextual embedding. However, there

are several reasons why MWE lexica remain relevant. One of the reasons is the inability of transformer-based models to understand phrase composition. Yu & Ettinger (2020) show that these models cannot understand the meaning of phrases. They rely on the words themselves for performance, which limits their ability to process complex language. Furthermore, Liu et al. (2024) find that idiomatic and compositionality-based tasks, such as idiom identification and interpretation, still fall behind human performance and require more task-specific optimization.

While MWEs are challenging for the LLMs of well-resourced languages, the problems multiply in the case of low-resourced languages, where large corpora are not available for training LLMs. Without sufficient training data, models struggle to learn the meanings of MWEs. Efforts such as the PARSEME shared tasks have made progress in providing datasets and tools for MWE identification in low-resource languages. However, there is still a need for more robust methods that can handle these expressions in scenarios where training data is limited.

Several studies highlight how lexica directly improve model performance for MWEs. Dimakis et al. (2024) showed that LLM scores in MWE translation improved when prompting the LLM with bilingual lexica. However, the best results were obtained when bilingual pairs exemplifying the MWE and its translation were used to prompt the LLM. Ke et al. (n.d.) use sentiment lexica such as SentiWordNet during pre-training to create sentiment-aware representations, which, while not MWE-specific, illustrate the broader role of lexica in enriching language representations. Similarly, Dhananjaya et al. (2024) use lexica for intermediate fine-tuning tasks to improve the performance of multilingual language models in low-resource settings by leveraging linguistic knowledge from high-resource languages. In addition, Guo et al. (n.d.) examines how lexica that capture stylistic features can improve the adaptability of LLMs to different language styles through meta-tuning.

At the same time, other studies aim to improve non-compositional and idiomatic representations by relying on MWE usage data instead of lexica. Zeng & Bhat (2022) use fine-tuning through an adapter-based approach to improve the model's representation of idiomatic expressions. Instead of using a lexicon, the model is fine-tuned using idiomatic sentence data extracted from the MAGPIE corpus. The same corpus is used in Zeng et al. (2023) to develop IEKG (Idiomatic Expression Knowledge Graph), a resource that uses knowledge graphs to capture the idiomatic usage of expressions. Unlike traditional lexica, IEKG organizes idioms into a structured framework that encodes detailed semantic relationships such as causes, effects, and attributes, and is specially designed to be used with

LLMs. MWEs in the CoAM dataset proposed by Ide et al. (2024) are defined as idiomatic sequences that satisfy three conditions: they consist of at least two fixed lexemes, display semantic, lexical, or syntactic idiomaticity, and are not proper nouns. The authors showed that fine-tuning with CoAM was highly effective in detecting MWEs and achieved higher F1 scores compared to previous lexicon-based methods. However, challenges remain, particularly in improving recall for unseen MWEs, highlighting the need for further research. In addition, a study by He et al. (2024) highlights the importance of advanced training strategies for improving idiomatic representations. Their approach uses dynamic triplets that include idiomatic expressions, correct paraphrases, and incorrect paraphrases. This allows models to learn the nuanced distinctions between idiomatic and literal meanings.

By contrast, some studies use LLMs to generate resources. Li et al. (2024) exemplify this shift by constructing IDIOMKB, a multilingual idiom knowledge base generated with LLMs. IDIOMKB encodes the figurative meanings of MWEs across languages, improving idiomatic translation without relying on traditional lexica. Such methods highlight the potential of LLMs to replace static lexica in specific contexts, providing flexible solutions to MWE-related tasks.

In summary, LLMs continue to exhibit notable limitations in the processing of MWEs, particularly in low-resourced contexts. Nevertheless, it remains premature to draw definitive conclusions regarding the specific conditions under which MWE resources most effectively contribute to the performance of LLMs. A comprehensive examination of this issue lies beyond the scope of the present survey. At the same time, it should be noted that certain aspects of lexicographic documentation, such as informal definitions and usage examples, also show promise with downstream tasks in low-resource settings.

A new research paradigm is emerging, indicating the necessity for MWE lexical databases to (re)define their contents, structure and role.

8 Discussion

We set out for this survey by examining several features of existing MWE resources and trying to identify the best encoding practices of useful information about MWEs. We worked on two tiers. On the first tier, we described the macro-properties of computational lexica such as linguality, availability, acquisition method, and linkage to external general lexica. On the second tier, we delved into the details of MWE documentation such as the forms that represent the MWE, morphosyntactic and semantic information, and availability of usage examples mainly through linking the MWE entry with their corpora occurrences.

In this section, we summarise and discuss our findings. We also add some remarks as a contribution to the discussion about recommendations on the development of MWE resources.

8.1 A bird's eye view of the MWE resource landscape

8.1.1 Macroscopic properties of the resources

Regarding linguality, most lexica are monolingual (e.g., Arabic, Portuguese, English) or bilingual (e.g., English-Spanish, Polish-English). The languages represented are predominantly European, with few exceptions such as Arabic, Chinese, Japanese, Hindi, Marathi, Korean, and Persian (as well as variants of Spanish spoken in Mexico, Argentina, Chile etc.). Less-resourced languages are underrepresented. The observation made by Losnegaard et al. (2016), namely that bilingual and multilingual MWE resources, including lexical ones, are hard to find, is still valid. English remains the best-represented language, appearing in nearly half of the language pairs as a source, target, or pivot language. The scarcity of bilingual and multilingual MWE lexica remains a significant challenge that could impede research and development of machine translation and other cross-lingual NLP tasks.

Most resources are MWE-dedicated, but some of them feature both one-word and multi-word entries. MWE dedicated resources are, in principle, independent and not linked to general lexica. Resources tend to address the general language with a few exceptions, such as expressions denoting sentiments. Again, a phenomenon of disregard for within-language diversity is observed, as these (predominantly colloquial) language aspects have not been well-documented.

On the availability front, it is good news that most of the resources are included in comprehensive international catalogs or language repositories and are freely available, at least for research purposes. On the other hand, the description of the contents of the resources often lacks the detail and clarity required to understand exactly what type of information the resources offer.

About half of the resources were developed (semi-)automatically and one-third of them manually. The literature does not offer benchmarks or diagnostics for measuring the quality of resources, whether constructed automatically or manually.

The size of the resources varies, but, in general, they are not extensive; this fact might reflect the effort required to develop such resources and the little motivation supporting their development.

A boost in their development is noticeable, however, in correlation with three major European projects: two COST actions (PARSEME and UniDive) and one Horizon project (ELEXIS): see Figure 1.

8.1.2 Details of MWE documentation

Three types of information could be stipulated as useful for both the human user and the state-of-the-art NLP: the MWE identifier (MWEI), the documentation of the MWE with usage examples, and the description of the idiomatic meaning of the MWE. Not all resources offer this information, while a small number of resources contribute significantly richer material.

The MWE identifier. The identifying representation of a MWE (MWEI) in a lexicon is the key to accessing the bundle of encoded MWE properties. In a way, the MWEI functions as an interface between the information about the MWE and its usages and has to be formulated accordingly. However, it should be kept in mind that computational MWE lexica are expensive resources and have to be adaptable to all conceivable applications, including those that involve the human usage of the lexicon. Of course, on the condition that the required information is provided, a type of MWEI can be converted into another, but only one identifying representation is normally chosen in each lexicon. The resources we examined can be divided into those that prioritise the needs of NLP processors and those that address both the NLP machines and the human user; the latter type of resources outnumbers the former significantly.

Naturally, the MWE lexicon that converges with NLP processors provides MWEIs that are usable by a particular technology. For instance, if the aim is to identify MWEs in a corpus, it may be advisable that the identifier adapt to the type of annotation provided. In such a case, the MWEI may be fully lemmatized and annotated with morphosyntactic information encoded in a particular formalism, the one that is used in the intended corpus (Leseva et al. 2024). This type of MWEI is not appropriate for humans who want to immediately recognize the MWE when they look it up in the lexicon.

Most resources adopt a type of MWEI that we have dubbed "the Lemma MWEI" that is aimed to satisfy both the human and the NLP needs. We found that there is some convergence regarding the definition of the Lemma MWEI: the fixed lexicalised MWE parts are not lemmatized, the inflecting lexicalised parts are lemmatized and the nominal arguments, if any, are rendered by indefinite pronouns. Simple as it may seem, several open issues still beg for clarification and

or/convergence; among others, those are the lemmatization strategies, the representation of the free complements, especially the clausal ones, and the number of free complements that should be codified. Variation, which is a wide-spread but not well-described phenomenon with MWEs cross-linguistically, also motivates a variety of approaches with an impact on the definition of the Lemma MWEI; whether variation should be encoded in the Lemma MWEI and to which extent is an open issue itself. Overall, there is ample room for discussion about further conversion between resources regarding the type of MWEI that best serves their purposes.

The documentation of the MWE with usage examples. Humans appreciate lexica that provide usage examples, but symbolic NLP does not use them. On the other hand, research in deep learning and LLM-based NLP reports significant benefits from the exploitation of usage examples. In order to provide this type of data, a MWE lexicon may be linked to a corpus. If a suitable corpus is not available, a number of usage examples can be elicited from independent resources, such as the Web, or even constructed manually and be provided as part of the documentation of the MWE, eventually creating a MWE-dense corpus linked to a lexicon. Suitable corpora are not always available for all types of MWEs, many of which "live" in colloquial language, which is not often represented in the available corpora.

Overall, we have found that a significant number of lexica are not linked to specific corpora, especially smaller or manually curated lexica.

The description of the idiomatic meaning of the MWE. Taking advantage of already aligned resources (such as wordnets, framenets) in which MWEs are identified and annotated as such is one rather fast way of creating a MWE-aware resource that inherits a lot of lexico-semantic information about the respective MWEs. The idiomatic (non-compositional) meaning is arguably the most noticeable characteristic of MWEs, but a description of it is provided by less than half of the examined resources. In these cases, the definition is inherited from general lexica or composed for this purpose as an informal one or a paraphrase. Furthermore, the MWE may be integrated in a network of semantic relations such as inheritance relations, lexico-semantic relations such as equivalence (synonymy) and antonymy relations. Only one resource in our list assigns embeddings to MWEs.

8.2 Towards defining minimum requirements and recommendations for MWE lexical resource development

Having sketched out the contemporary MWE resource landscape, we now turn to the question of which are minimum requirements for effective MWE resource development. Of course, lexica – like any other type of language resources – are developed to aid or support specific applications and usages. However, some fundamental requirements ensure their interoperability, scalability, and (re-)usability across different NLP tasks and applications. We argue that these core requirements might help maintain a minimum level of consistency in representation and integration with computational frameworks, regardless of the specific application they were intended for.

Symbolic NLP certainly requires MWE lexica. Deep learning and transformer-based approaches to NLP may use lexical information about MWEs, in particular in the case of less-resourced languages (Savary et al. 2019a, Section 7). With MWE lexica the question seems to be which information encoded in which way would be appropriate for which technologies. And, of course, which would be the required macroscopic properties of MWE lexica.

Our survey has shown that three types of information are used by state-of-the-art NLP (supervised approaches and LLMs); these findings could probably be useful in designing MWE lexical resources and MWE-aware LLM benchmarks:

- **the MWE identifier (MWEI)**: it plays the role of an interface between humans and lexica and between lexica and the NLP technology that uses them; often the MWEI itself is the required information. The Lemma MWEI seems to be more suitable for all these functions, and its exact form has to adjust to the requirements of the technologies to which it is addressed.

- **usage examples**: they can be used by all types of state-of-the-art NLP and can be obtained through linking lexica to specific corpora. Alternatively, collections of usage examples can be developed for each MWE (indicatively MAGPIE (Haagsma et al. 2020), COAM (Ide et al. 2024)). One problem is that MWEs are mostly used in colloquial speech, which is relatively underrepresented in the available corpora. It would be advisable that both corpus linking and collections of usage examples follow guidelines that would ensure the best representation of the forms that the MWE can assume and the contexts where it is used (Ide et al. 2024, Zeng et al. 2023). Of course, there is an acute problem of sparsity of data for low-resourced languages.

- **meaning definition**: in textual form, it is necessary for human users, but it seems that also LLMs can take advantage of it when used for downstream

applications. If definitions can be proven to robustly facilitate downstream applications of LLMs, their further development would require first establishing a set of best practices.[24]

In addition, recommendations are necessary regarding the macroscopic properties of lexica. Building on Arranz et al. (2008), we make four recommendations that naturally come out of the analysis of the landscape of MWE lexica throughout this chapter. We also add a new recommendation, given that the systems need to cope with applications such as sentiment analysis.

1. Document the design and the contents of the resource thoroughly, clearly, and concisely. This ensures its usage in appropriate tasks to its full potential.

2. Ensure that the resource is easily accessible through stable and friendly repositories. Cataloging and advertising the resource on internationally recognized and carefully maintained platforms ensures its long-term visibility.

3. Make the resource free at least for research purposes. A license attached to the resource clarifies its usage possibilities and facilitates its effective usage, while the absence of a license can can negatively impact its usage, running contrary to the purposes of its development.

4. Ensure maintenance of the resource over time. Given the scarcity of such linguistic descriptions, not maintaining existing ones can worsen the situation.

5. Cover special usages of language as well, such as expressions denoting sentiments. These are less familiar, even to native speakers, and can lead to misinterpretations of texts containing them.

9 Conclusion and outlook for future research

Dealing with the automatic identification of MWEs in text, their semantic non-compositionality, and distributional properties is still challenging for NLP. MWE-dedicated lexica are essential resources to address these challenges (Savary et al. 2019a). In this chapter, we presented the results of a survey on MWE lexica,

[24]Instructions for good data definitions addressed to humans exist, for instance https://businessintelligence.gwu.edu/creating-good-data-definitions.

which we conducted as a follow-up to the survey by Losnegaard et al. (2016). A total of 66 resources, both MWE-aware and MWE-dedicated, all developed in the years 2016 to 2024, were identified through collaborative labor. We conducted a literature survey and a survey sent to the community.

The primary objective of this survey was to offer a comprehensive overview of the MWE-related computational lexica. As mentioned, we have identified 66 resources using strict inclusion criteria, with special emphasis on the languages featured in the resources and their level of representation. Thus, our endeavour is following the UniDive COST Action idea to highlight the extent to which less-represented languages are included in existing resources. In addition to cataloguing the MWE lexica, we investigated the strategies the developers used to encode lexical information for MWEs, at the levels of lemma, semantics, syntax, as well as linking to corpora. One of the objectives was to touch upon the issue of the role of MWE lexical resources in the era of LLMs. Addressing all these objectives should strengthen the efforts to enhance the development of MWE lexica by defining minimum requirements and recommendations.

Future work should focus on the formats developers use to encode lexical information in MWE lexica and the extent to which these resources are compatible with - or can be mapped onto - existing standards. Moreover, while existing standards provide structured representations, challenges remain in ensuring that diverse MWE-related phenomena are captured in a way that both universality and diversity are ensured. Research on MWE lexica should look into existing standardized vocabularies.

Processing non-compositional expressions and complex language still presents a challenge for LLMs and falls behind human performance. Problems exist for well-represented languages, and for low-represented ones, they are multiplied. The PARSEME shared tasks provided some datasets and tools for the identification of MWE for low-resourced languages, but they still need more training data. Some studies have shown that lexica, knowledge graphs and contrastive learning improve model performance for MWEs. An important avenue for further research is the interplay between LLMs and the use of MWE-related lexica for downstream NLP tasks.

Also promising with downstream tasks for low-resourced languages are some aspects of lexicographic documentation – informal definitions and usage examples. As multimodal NLP continues to advance, a new avenue of research is opening up—one that calls for the systematic treatment of MWEs in modalities beyond text, including speech, gesture, and image captions.

Although our focus in this article is on NLP-oriented lexica, we believe that the human user should not be neglected in the design of a MWE lexicon, in particular

in the case of less-resourced languages: a MWE lexicon is an expensive resource and will be created only once, if at all. In this regard, it should be essential to explore and quantify the usability of such information.

Based on the recommendations and looking ahead, research on MWE lexica should prioritize the proof-of-concept lexical encoding of MWEs. It seems that the expertise of the lexicographer is still useful in the LLM era. The lexicographer has to work both intensionally (provide definitions) and extensionally (provide usage examples), especially when documenting an under-resourced language and redefine her methods in the new technological landscape.

Limitations

In this survey, we have focused on a selection of MWE resources based on a number of criteria in terms of their coverage and the level of linguistic description (see Section 3). We have included resources for which we have found sufficient information in the sources that we surveyed; in addition, we have collected information about MWE resources from the academic community actively involved in MWE research.

However, we recognize that our efforts have certain limitations, as we primarily covered bibliographic sources in English. We acknowledge that there are large research communities that use other languages, such as Spanish, French, Russian, Arabic, Chinese, etc., which are not accessible to us and thus not included in the present survey. In addition, the survey focuses on resources that encode the idiomatic aspects of MWEs despite the fact that the same structures often have compositional usages along with idiomatic ones (Savary et al. 2019b).

Abbreviations

ACL	Association for Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
ELG	European Language Grid
LLM	Large Language Model
MWE	Multiword Expression
MWEI	Multiword Expression Identifier
NLP	Natural Language Processing
PARSEME	Parsing and Multi-word Expressions
POS	Part of Speech

Acknowledgements

This work has received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Anders, Valentin. 2022a. *Chilenismos (deChile)*. https://live.european-language-grid.eu/catalogue/lcr/12730.

Anders, Valentin. 2022b. *Expressions (deChile)*. https://live.european-language-grid.eu/catalogue/lcr/12730.

Antunes, Sandra & Amália Mendes. N.d. MWE in Portuguese: Proposal for a Typology for Annotation in Running Text. In Valia Kordoni, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the 9th Workshop on Multiword Expressions*, 87–92. Atlanta: Association for Computational Linguistics. https://aclanthology.org/W13-1013/.

Arranz, Victoria, Franck Gandcher, Valérie Mapelli & Khalid Choukri. 2008. A guide for the production of reusable language resources. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech: European Language Resources Association (ELRA). https://aclanthology.org/L08-1434/.

Al-Badrashiny, Mohamed. 2016. SAMER: A semi-automatically created lexical resource for Arabic verbal multiword expressions tokens paradigm and their morphosyntactic features. In Koiti Hasida (ed.), *Proceedings of the 12th Workshop on Asian Language Resources*, 113–122. Osaka: The COLING 2016 Organizing Committee. https://aclanthology.org/W16-5414.

Baker, Collin F., Charles J. Fillmore & John B. Lowe. N.d. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 86–90. Montreal: Association for Computational Linguistics. DOI: 10.3115/980845.980860. https://aclanthology.org/P98-1013/.

Barančíková, Petra & Václava Kettnerová. 2017. ParaDi: Dictionary of paraphrases of Czech complex predicates with light verbs. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 1–10. Valencia: Association for Computational Linguistics. DOI: 10.18653/v1/W17-1701.

Barbu Mititelu, Verginica, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stanković & Christian Chiarcos. 2024. Multiword expressions between the corpus and the lexicon: Universality, idiosyncrasy, and the lexicon-corpus interface. In Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre & Alexandre Rademaker (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies*, 147–153. Torino: ELRA & ICCL. https://aclanthology.org/2024.mwe-1.19.

Barbu Mititelu, Verginica, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova & Maria Todorova. 2019. Hear about verbal multiword expressions in the Bulgarian and the Romanian Wordnets straight from the horse's mouth. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 2–12. Sofia: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5102.

Beaugrande, Robert De & Wolfgang U. Dressler. 1981. *Introduction to text linguistics*. London: Longman.

Bejček, Eduard. 2017. *Czech verbal MWEs*. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2603.

El-Beltagy, Samhaa R. 2016. NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2900–2905. Portorož: European Language Resources Association (ELRA). https://aclanthology.org/L16-1463.

Bhagat, Rahul & Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics* 39(3). 463–472. DOI: 10.1162/COLI_a_00166. https://aclanthology.org/J13-3001.

Bielinskienė, Agnė, Loïc Boizou, Ieva Bumbulienė, Jolanta Kovalevskaitė, Tomas Krilavičius, Justina Mandravickaitė, Erika Rimkutė, Jurgita Vaičenonienė & Laura Vilkaitė-Lozdienė. 2022. The database of Lithuanian multiword expressions. *https://arka.pastovu.vdu.lt/*. https://clarin.vdu.lt/xmlui/handle/20.500.11821/49.

Blagus Bartolec, Goranka, Gorana Duplančić Rogošić & Antonia Ordulj. 2024. INIKOL - Collocational Database for Learning Croatian as a Foreign Language. In Amy Qiu, Bill Noble, David Pagmar, Vladislav Maraev & Nikolai Ilinykh (eds.), *Proceedings of the 2024 CLASP Conference on Multimodality and Interac-*

tion in Language Learning, 8–12. Gothenburg: Association for Computational Linguistics. https://aclanthology.org/2024.clasp-1.2/.

Bogantes, Diana, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez & Agata Savary. 2016. Towards lexical encoding of multi-word expressions in Spanish dialects. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of LREC*, 2255–2261. Portorož: European Language Resources Association (ELRA). https://aclanthology.org/L16-1358/.

Bulkes, Nyssa Z. & Darren Tanner. 2017. "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behaviour Research Methods* 49(2). 772–783. DOI: 10.3758/s13428-016-0747-8.

Czerepowicka, Monika & Agata Savary. 2015. SEJF -a Grammatical Lexicon of Polish Multi-Word Expressions. In *Proceedings of the Language Technology Conference 2015 (LTC 2015)*, 5. Poznań. https://hal.science/hal-01223683.

Dhananjaya, Vinura, Surangika Ranathunga & Sanath Jayasena. 2024. Lexicon-based fine-tuning of multilingual language models for low-resource language sentiment analysis. *CAAI Transactions on Intelligence Technology* 9(5). 1116–1125. DOI: 10.1049/cit2.12333.

Dimakis, Antonios, Stella Markantonatou & Antonios Anastasopoulos. 2024. Dictionary-aided translation for handling multi-word expressions in low-resource languages. In Lun-Wei Ku, Andre Martins & Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, 2588–2595. Bangkok: Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.152.

Dutch Language Institute. 2016. *Referentiebestand Belgisch-Nederlands*. European Language Grid. https://live.european-language-grid.eu/catalogue/lcr/13397.

ELRA. 2010. *Terminology database of expressions*. https://live.european-language-grid.eu/catalogue/lcr/2853.

ELRA. 2019. *English-Persian database of idioms and expressions*. https://live.european-language-grid.eu/catalogue/lcr/2863.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge: The MIT Press.

Filipović Petrović, Ivana, Miguel López Otal & Slobodan Beliga. 2024. Croatian idioms integration: Enhancing the LIdioms multilingual linked idioms dataset. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*

(LREC-COLING 2024), 4106–4112. Torino: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.366/.

Fisas, Beatriz, Luis Espinosa-Anke, Joan Codina-Filbá & Leo Wanner. 2020. CollFrEn: Rich bilingual English–French collocation resource. In Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova & Agata Savary (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 1–12. Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.1.

Fran Ramovš Institute of the Slovenian Language ZRC SAZU. N.d. *Terminological multiword expressions lexicon.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1780.

Gaspari, Federico, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis & Andy Way. 2023. Digital language equality: Definition, metric, dashboard. In Georg Rehm & Andy Way (eds.), *European Language Equality: A Strategic Agenda for Digital Language Equality*, 39–73. Cham: Springer International Publishing. DOI: 10.1007/978-3-031-28819-7_3.

Giouli, Voula. 2023. A model for representing the semantics of MWEs: From lexical semantics to the semantic annotation of complex predicates. *Frontiers in Artificial Intelligence* 6. DOI: 10.3389/frai.2023.802218.

Giouli, Voula, Vera Pilitsidou & Hephaestion Christopoulos. 2020. Greek within the Global FrameNet initiative: Challenges and conclusions so far. In Tiago T. Torrent, Collin F. Baker, Oliver Czulo, Kyoko Ohara & Miriam R. L. Petruck (eds.), *Proceedings of the International FrameNet Workshop 2020: Towards a global, multilingual FrameNet*, 48–55. Marseille: European Language Resources Association. https://aclanthology.org/2020.framenet-1.7/.

Giouli, Voula, Vera Pilitsidou & Hephestion Christopoulos. 2024. A FrameNet approach to deep semantics for MWEs. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 147–186. Berlin: Language Science Press.

Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1/2). 23–39. DOI: 10.1007/s10579-009-9094-z.

Guo, Ruohao, Wei Xu & Alan Ritter. N.d. Meta-Tuning LLMs to Leverage Lexical Knowledge for Generalizable Language Style Understanding. In Lun-Wei Ku, Andre Martins & Vivek Srikumar (eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 13708–13731. Bangkok: Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.740. https://aclanthology.org/2024.acl-long.740/.

Haagsma, Hessel, Johan Bos & Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 279–287. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.35/.

He, Wei, Marco Idiart, Carolina Scarton & Aline Villavicencio. 2024. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In Lun-Wei Ku, Andre Martins & Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, 12473–12485. Bangkok: Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.741.

Hedegaard, Jette & Thomas Troelsgård. 2010. Dice in the Web: An online Spanish collocation dictionary. In Sylviane Granger & Magali Paquot (eds.), *eLexicography in the 21st century: new challenges, new applications. Proceedings of ELEX2009, Louvain-la-Neuve, 22-24 October 2009*, 369–374. Louvain-la-Neuve: Cahiers Du Cental 7. Presses Universitaires de Louvain. https://api.semanticscholar.org/CorpusID:2469754.

Hubers, Ferdy, Catia Cucchiarini & Helmer Strik. 2020. Dedicated language resources for interdisciplinary research on multiword expressions: Best thing since sliced bread. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4418–4425. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.544.

Ide, Yusuke, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito & Taro Watanabe. 2024. *CoAM: Corpus of All-Type Multiword Expressions*. https://arxiv.org/abs/2412.18151.

Institute of the Estonian Language. 2016. *The Dictionary of Estonian Synonyms*. European Language Grid. https://live.european-language-grid.eu/catalogue/lcr/13590.

Iñurrieta, Uxoa, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka & Kepa Sarasola. 2018. Konbitzul: An MWE-specific database for Spanish-Basque. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Re-*

sources and Evaluation (LREC 2018). Miyazaki: European Language Resources Association (ELRA). https://aclanthology.org/L18-1397/.

Jochim, Charles, Francesca Bonin, Roy Bar-Haim & Noam Slonim. N.d. SLIDE - a Sentiment Lexicon of Common Idioms. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA). https://aclanthology.org/L18-1379/.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali & Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 6282–6293. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.560.

Kanzaki, Kyoko & Hitoshi Isahara. 2019. Towards linking synonymous expressions of compound verbs to Japanese WordNet. In Piek Vossen & Christiane Fellbaum (eds.), *Proceedings of the 10th Global Wordnet Conference*, 185–190. Wroclaw: Global Wordnet Association. https://aclanthology.org/2019.gwc-1.23.

Ke, Pei, Haozhe Ji, Siyang Liu, Xiaoyan Zhu & Minlie Huang. N.d. SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge. In Bonnie Webber, Trevor Cohn, Yulan He & Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6975–6988. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.567. https://aclanthology.org/2020.emnlp-main.567/.

Khokhlova, Maria. 2020. Collocations in Russian Lexicography and Russian Collocations Database. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3198–3206. European Language Resources Association. https://aclanthology.org/2020.lrec-1.391.

Koeva, Svetla, Ivelina Stoyanova, Maria Todorova & Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of Bulgarian multiword expressions. In *Proceedings of the Workshop on Lexicographic Resources for Human*

Language Technology, 86–95. https://ailab.ijs.si/globalex/files/2016/06/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf#page=92.

Krek, Simon, Apolonija Gantar, Cyprian Laskowski, Luka Krsnik, Iztok Kosem, Janez Brank, Kaja Dobrovoljc, Špela Arhar Holdt, Jaka Čibej, Marko Robnik-Šikonja, Bojan Klemenc & Vojko Gorjanc. 2021. Multiword Expressions lexicon extracted from the Gigafida 2.1 corpus. *http://slovnica.ijs.si/*. https://www.clarin.si/repository/xmlui/handle/11356/1421.

Krstev, Cvetana, Ivan Obradović, Ranka Stanković & Duško Vitas. 2013. An approach to efficient processing of multi-word units. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem & Piotr Fuglewicz (eds.), *Computational Linguistics*, vol. 458, 109–129. Series Title: Studies in Computational Intelligence. Berlin: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-34399-5_6.

Kurfalı, Murathan, Robert Östling, Johan Sjons & Mats Wirén. 2020. A multiword expression dataset for Swedish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4402–4409. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.542.

Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press.

Li, Shuang, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao & Yanghua Xiao. 2024. Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38(17). 18554–18563. DOI: 10.1609/aaai.v38i17.29817.

Lichte, Tim, 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 multi-word expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.

Liebeskind, Chaya & Yaakov HaCohen-Kerner. 2016. A lexical resource of Hebrew verb-noun multi-word expressions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources*

and Evaluation (LREC'16), 522–527. Portorož: European Language Resources Association (ELRA). https://aclanthology.org/L16-1083/.

Liebeskind, Chaya & Yaakov HaCohen-Kerner. 2018. Verbal multi-word expressions in Yiddish. In Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais & Farid Meziane (eds.), *Natural language processing and information systems*, 205–216. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-91947-8_20.

Liebeskind, Chaya & Yaakov HaCohen-Kerner. N.d. Semantically Motivated Hebrew Verb-Noun Multi-word Expressions Identification. In Yuji Matsumoto & Rashmi Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1242–1253. Osaka: The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1118/.

Liu, Yang, Melissa Xiaohui Qin, Hongming Li & Chao Huang. 2024. Revisiting a pain in the neck: Semantic phrase processing benchmark for language models. *arXiv preprint arXiv:2405.02861*.

Ljubešić, Nikola, Kaja Dobrovoljc & Darja Fišer. 2015. *MWELex – MWE lexica of Croatian, Slovene and Serbian extracted from parsed corpora. *Informatica* 39(3). https://www.informatica.si/index.php/informatica/article/view/985.

Ljubešić, Nikola, Kaja Dobrovoljc, Simon Krek, Marina Peršuric Antonic & Darja Fišer. 2014. hrMWELex – a MWE lexicon of Croatian extracted from a parsed gigacorpus. In *9th Language Technologies Conference Information Society (IS 2014)*, 25–31.

Lobzhanidze, Irina. 2019. Computational model of the modern Georgian language and search patterns for an online dictionary of idioms. In Alexandra Silva, Sam Staton, Peter Sutton & Carla Umbach (eds.), *Language, Logic, and Computation. TbiLLC 2018*, vol. 11456 (Lecture Notes in Computer Science), 187–208. Berlin: Springer.

Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann & Johanna Monti. 2016. PARSEME survey on MWE resources. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2299–2306. Portorož: European Language Resources Association (ELRA). https://aclanthology.org/L16-1364/.

Markantonatou, Stella, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chirila, Dimitrios Karamatskos, Nicolaos Valeontisa & George Pavlidis. 2024. Description of Pomak within IDION: Challenges in the representation of verb multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Mul-*

tiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives. Berlin: Language Science Press. DOI: 10 . 5281 / zenodo . 10998633.

Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 130–134. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5115. https://aclanthology.org/W19-5115.

Masini, Francesca, M. Silvia Micheli, Andrea Zaninello, Sara Castagnoli & Malvina Nissim. 2020. *MWE_combinet_release_1.0*. Associazione Italiana di Linguistica Computazionale. https://amsacta.unibo.it/id/eprint/6506/.

Maynard, Diana, Joanna Wright, Mark A. Greenwood & Kalina Bontcheva. 2022. *D1.11 Report on the English Language*. Tech. rep. European Language Equality. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_11__Language_Report_English_.pdf.

Maziarz, Marek, Łukasz Grabowski, Tadeusz Piotrowski, Ewa Rudnicka & Maciej Piasecki. 2023. Lexicalisation of Polish and English word combinations: An empirical study. *Poznan Studies in Contemporary Linguistics* 59(2). 381–406. DOI: 10.1515/psicl-2023-2002.

Mel'čuk, Igor A. 2006. Explanatory combinatorial dictionary. In Giandomenico Sica (ed.), *Open problems in linguistic and lexicography*, 225–355. Monza: Polimetrica.

Mel'čuk, Igor A. 2012. *Semantics: From meaning to text*, vol. I. Amsterdam: John Benjamins.

Miletić, Filip & Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics* 12. 593–612. DOI: 10.1162/tacl_a_00657.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.

Moussallem, Diego. 2018. LIdioms: A Multilingual Linked Idioms Data Set. In Nicoletta alzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1392.

Navigli, Roberto & Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193. 217–250. DOI: 10.1016/j.artint.2012.07.001.

Nevěřilová, Zuzana. 2018. Discovering continuous multi-word expressions in Czech. *Computación y Sistemas* 22(3). DOI: 10.13053/cys-22-3-3022.

Odijk, Jan & Martin Kroon. N.d. A Canonical Form for Flexible Multiword Expressions. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 91–101. Torino: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.8/.

Õim, Asta. 2000. *Fraseoloogiasõnaraamat*. 2-ne, täiendatud ja parandatud trükk. Tallinn: Eesti keele sihtasutus. https://arhiiv.eki.ee/dict/frs/frs.html.

Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.

Pagliai, Irene. 2023. Bridging the gap: Creation of a lexicon of 150 pairs of English and Italian idioms including normed variables for the exploration of idiomatic ambiguity. *Journal of Open Humanities Data* 9. 16. DOI: 10.5334/johd.123.

Pecina, Pavel. 2008. *Gold standard reference data for multiword expression extraction: Czech dependency bigrams from the Prague Dependency Treebank*. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1457.

Phelps, Dylan, Thomas Pickard, Maggie Mi, Edward Gow-Smith & Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre & Alexandre Rademaker (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, 178–187. Torino: ELRA & ICCL. https://aclanthology.org/2024.mwe-1.22.

Piperidis, Stelios. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. 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 2012)*, 36–42. Istanbul: European Language Resources Association (ELRA). https://aclanthology.org/L12-1647/.

Puzyrev, Dmitry, Artem Shelmanov, Alexander Panchenko & Ekaterina Artemova. 2019. A dataset for noun compositionality detection for a Slavic language. In Tomaž Erjavec, Michał Marcińczuk, Preslav Nakov, Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger & Roman Yangarber (eds.), *Proceedings of the 7th workshop on balto-slavic natural language processing*, 56–62. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/W19-3708.

Ramisch, Carlos. 2016. How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality. In Katrin Erk (ed.), *How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality*, 156–161. Berlin: Association for Computational Linguistics. DOI: 10.18653/v1/P16-2026.

Ramisch, Carlos, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh & Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova & Agata Savary (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 107–118. Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.14/.

Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In Haifeng Wang & David Yarowsky (eds.), *Proceedings of 5th International Joint Conference on Natural Language Processing*, 210–218. Chiang Mai: Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1024.

Rehm, Georg. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe* (Cognitive Technologies). Springer. Cham.

Robertson, Frankie. 2020. Filling the ___-s in Finnish MWE lexicons. In Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova & Agata Savary (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 13–21. Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.2.

Rodríguez, Adolfo Enrique. 2022. *Lunfardo Dictionary*. European Language Grid. https://live.european-language-grid.eu/catalogue/lcr/12494.

Rodriguez, Barrios & Maria Auxiliadora. 2019. A Spanish E-dictionary of collocations. In Kim Gerdes (ed.), *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 160–167. DOI: 10.18653/v1/W19-7719.

Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 2019a. Without lexicons, multiword expression identification will never fly: A position statement. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5110.

Savary, Agata & Silvio Ricardo Cordeiro. 2017. Literal readings of multiword expressions: As scarce as hen's teeth. In Jan Hajič (ed.), *Literal readings of multiword expressions: As scarce as hen's teeth*, 64–72. Prague. https://aclanthology.org/W17-7610.

Savary, Agata, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019b. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics* 112. 5–54. https://ufal.mff.cuni.cz/pbml/112/art-savary-et-al.pdf.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 31–47. Valencia: Association for Computational Linguistics. DOI: 10.18653/v1/W17-1704.

Savary, Agata, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz & Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In Maite Melero, Sakriani Sakti & Claudia Soria (eds.), *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, 372–382. Torino: ELRA & ICCL. https://aclanthology.org/2024.sigul-1.45.

Schulte im Walde, Sabine. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998645. https://zenodo.org/doi/10.5281/zenodo.10998645.

Singh, Dhirendra, Sudha Bhingardive & Pushpak Bhattacharyya. 2016. Multiword Expressions Dataset for Indian Languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 2331–2335. Portorož: European Language Resources Association (ELRA). https://aclanthology.org/L16-1369/.

Skoumalová, Hana, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondřička & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press.

Todirascu, Amalia, Marion Cargill & Thomas Francois. N.d. PolylexFLE : Une base de données d'expressions polylexicales pour le FLE (PolylexFLE : A database of multiword expressions for French L2 language learning). In Emmanuel Morin, Sophie Rosset & Pierre Zweigenbaum (eds.), *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. volume i : Articles longs*, 143–156. Toulouse: ATALA. https://aclanthology.org/2019.jeptalnrecital-long.10/.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron & Fabre Lambeau. 2004. Lexical encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 80–87. Barcelona: Association for Computational Linguistics. https://aclanthology.org/W04-0411.

Volodina, Elena, David Alfter & Therese Lindström Tiedemann. 2024. Profiles for Swedish as a second language: Lexis, grammar, morphology. In Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg & Anna Blåder (eds.), *Proceedings of the Huminfra Conference (HiC 2024)*, 10–19. Gothenburg. DOI: 10.3384/ecp205002.

Vondřička, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112(1). 83–101. DOI: 10.2478/pralin-2019-0003.

Wada, Takashi, Yuji Matsumoto, Timothy Baldwin & Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. In Anna Rogers, Jordan Boyd-Graber & Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, 4732–4746. Toronto: Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-acl.290.

Walsh, Abigail, Teresa Lynn & Jennifer Foster. 2019. Ilfhocail: A lexicon of Irish MWEs. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović & Verginica Barbu Mititelu (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 162–168. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/W19-5120.

Wilkens, Rodrigo, Leonardo Zilio, Silvio Ricardo Cordeiro, Felipe Paula, Carlos Ramisch, Marco Idiart & Aline Villavicencio. 2017. LexSubNC: A dataset of lexical substitution for nominal compounds. In Claire Gardent & Christian Retoré (eds.), *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*. https://aclanthology.org/W17-6941/.

Yu, Lang & Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In Bonnie Webber, Trevor Cohn, Yulan He & Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4896–4907. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.397.

Zeng, Ziheng & Suma Bhat. 2022. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics* 10. 1120–1137. DOI: 10.1162/tacl_a_00510.

Zeng, Ziheng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou & Suma Bhat. 2023. IEKG: A commonsense knowledge graph for idiomatic expressions. In Houda Bouamor, Juan Pino & Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14243–14264. Singapore: Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.881.

Appendix

Table 2: List of resources included in the survey with basic reference information. Method: M – manual processing; S/A – semi-automatic; A – automatic. MWEI: Can – canonical lemma; Abs – abstract lemma (each component of the MWE is lemmatised). Meaning representation: GLOSS – gloss provided in the resource; EXTGLOSS – gloss from an external resource; STYL – stylistic information; PAR – paraphrases; SEMREL – semantic relations to other words (e.g., synonyms, etc.); POL – polarity information; COMPOS – compositionality ratings; SEMCLASS – semantic class; SEMFEAT – semantic features; COMPS – meaning of components; LEXF – lexical functions; VECT – vector representation. Un – Unknown.

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
Feature-NN (Schulte im Walde 2024)								
http://www.ims.uni-stuttgart.de/data/feature-comp-nn								
2025	DE	S/A	free	NO	Un	Un	COMPOS; GLOSS	Un
Bulgarian Integrated Lexicon (Osenova & Simov 2024)								
2024	BG	Un	Un	NO	YES	Un	EXTGLOSS	YES
Collocational Database for Learning Croatian as a Foreign Language (Blagus Bartolec et al. 2024)								
2024	HR, EN	Un	free	NO	NO	Un	NO	NO
MWE-CEFR Profiles (Volodina et al. 2024)								
https://spraakbanken.gu.se/larkalabb/svlp								
2024	SV	S/A	free	YES	NO	Un	EXTGLOSS	Un
MWEs in FrameNet-EL (Giouli et al. 2024)								
2024	EL	Un	Un	NO	Un	Can	EXTGLOSS	YES

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
DUCAME (Odijk & Kroon n.d.)								
https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP								
2023	NL	Un	Un	YES	Un	Can	GLOSS	YES
IDION POMAK (Markantonatou et al. 2024)								
https://pomak.idion.athenarc.gr								
2023	POMAK	M	free	YES	Un	Can	GLOSS; SEMRELS	Un
IdiomKB (Li et al. 2024)								
https://github.com/lishuang-w/IdiomKB								
2023	EN, ZH, JA	S/A	free	NO	NO	Can	EXTGLOSS	NO
IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions (Zeng et al. 2023)								
https://github.com/zzeng13/IEKG								
2023	EN	S/A	free	YES	YES	Can	EXTGLOSS	NO
Normed lexicon of English and Italian idioms (Pagliai 2023)								
https://data.goettingen-research-online.de/dataset.xhtml?persistentId=doi:10.25625/EPSWDY								
2023	EN, IT	Un	free for specific uses	NO	NO	Un	COMPOS	YES
Srp_DELAC (Krstev et al. 2013)								
https://llod.jerteh.rs/Srp_DELAC/								
2023	SR	M	academic	NO	NO	Can	SEMCLASS	YES

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
Terminological MWE lexicon (Fran Ramovš Institute of the Slovenian Language ZRC SAZU n.d.)								
https://live.european-language-grid.eu/catalogue/lcr/21521								
2023	SL	M	free	NO	NO	Can+Abs	NO	NO
Croatian dictionary of idioms (Filipović Petrović et al. 2024)								
https://lexonomy.elex.is/#/frazeoloskirjecnikhr/835								
2022	HR	S/A	free	YES	NO	Un	EXTGLOSS	NO
Expressions (deChile) (Anders 2022b)								
https://live.european-language-grid.eu/catalogue/lcr/12730/overview/								
2022	ES	M	free	NO	NO	Can	GLOSS	NO
Idioms of Chile [Chilenismos] (Anders 2022a)								
https://live.european-language-grid.eu/catalogue/lcr/14819								
2022	ES-CL	M	free	NO	NO	Can	GLOSS	NO
LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds (Wilkens et al. 2017)								
https://pageperso.lis-lab.fr/~carlos.ramisch/?page=downloads/compounds								
2022	PT	S/A	Un	NO	NO	Can	COMPS; GLOSS	Un
Lunfardo Dictionary (Rodríguez 2022)								
https://live.european-language-grid.eu/catalogue/lcr/12494								
2022	ES-AR	M	free	NO	NO	Can	GLOSS	NO
The Database of Lithuanian MWEs (Bielinskienė et al. 2022)								
https://resursai.pastovu.vdu.lt/paieska/paprastoji								

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
2022	LT	S/A	free	YES	YES	Can	SEMREL; SEM-CLASS	YES
Diretes: DIccionario RETicular de ESpañol (Rodriguez & Auxiliadora 2019)								
http://diretes.es/								
2021	ES	Un	Un	NO	YES	Can	LEXF; SEMRELS	YES
Lexicalisation of Polish and English word combinations (Maziarz et al. 2023)								
https://clarin-pl.eu/dspace/handle/11321/853								
2021	PL, EN	M	free	NO	NO	Can	PAR; COM-POS	NO
MWE lexicon extracted from the Gigafida 2.1 corpus (Krek et al. 2021)								
https://www.clarin.si/repository/xmlui/handle/11356/1421								
2021	SL	S/A	Un	YES	NO	Can	EXTGLOSS	YES
CollFrEn: Rich Bilingual English–French Collocation Resource (Fisas et al. 2020)								
https://github.com/TalnUPF/CollFrEn								
2020	EN, FR	A	free	YES	NO	Can	EXTGLOSS; SEMRELS; LEXF; VECT	YES
Dutch idiomatic expressions (Hubers et al. 2020)								
2020	NL	A	free	NO	NO	Can	GLOSS; COMPOS	Un
FinnMWE: a lexicon of Finnish MWEs (Robertson 2020)								
https://github.com/frankier/finnmwe								
2020	FI	A	free	NO	YES	Abs	NO	YES
LEMUR (Vondřička 2019)								
https://yggdrasil2.korpus.cz/ratatosk-paw/search/lemur								

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
2020	CZ	S/A	Un	NO	NO	Can	GLOSS; STYL	YES
MWE Dataset for Swedish (Kurfalı et al. 2020)								
https://github.com/MurathanKurfali/swedish-mwe-dataset								
2020	SV	S/A	free	NO	NO	Can	COMPOS	YES
MWE_combinet_release_1.0 (Masini et al. 2020)								
https://amsacta.unibo.it/id/eprint/6506/								
2020	IT	S/A	free	YES	NO	Abs	NO	YES
English-Persian database of idioms and expressions (ELRA 2019)								
https://live.european-language-grid.eu/catalogue/lcr/2863								
2019	EN, FA	S/A	paid	NO	YES	Can	NO	Un
IDION: A database for Modern Greek MWEs (Markantonatou et al. 2019)								
2019	EL	Un	free	YES	YES	Can	SEMRELS; GLOSS	YES
ilFhocail (Walsh et al. 2019)								
https://live.european-language-grid.eu/catalogue/lcr/13464								
2019	GA	M	Un	NO	NO	Can	NO	NO
MWE dictionary for Bulgarian and Romanian (Barbu Mititelu et al. 2019)								
2019	BG, RO	S/A	free	NO	YES	Can	EXTGLOSS; PAR	YES
Noun Compound Dataset for Russian (Puzyrev et al. 2019)								
https://github.com/s-nlp/ru-comps								
2019	RU	S/A	free	NO	NO	Can	COMPOS	YES
PolylexFLE (Todirascu et al. n.d.)								
https://github.com/amaliatodirascu/PolylexFLE								

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
2019	FR	Un	Un	NO	YES	Un	GLOSS	YES
Japanese compound verb lexicon (Kanzaki & Isahara 2019)								
https://vvlexicon.ninjal.ac.jp/en/								
2018	JA, EN, ZH, KO	Un	free	NO	NO	Can	SEMFEAT; COMPS	YES
Konbitzul (Iñurrieta et al. 2018)								
https://live.european-language-grid.eu/catalogue/lcr/13707								
2018	EU, ES	M	free	NO	NO	Can	NO	NO
LIDIOMS: A Multilingual Linked Idioms Data Set (Moussallem 2018)								
https://github.com/dice-group/LIdioms								
2018	EN, DE, IT, PT, RU	S/A	free	NO	YES	Can	GLOSS	Un
Polish verbal MWEs (Savary & Cordeiro 2017)								
http://clip.ipipan.waw.pl/MweLitRead								
2018	PL	A	free	YES	NO	Un	SEMCLASS	YES
SLIDE: Sentiment Lexicon of IDiomatic Expressions (Jochim et al. n.d.)								
https://research.ibm.com/haifa/dept/vst/debating_data.shtml								
2018	EN	S/A	free	NO	YES	Can	POL	NO
Verbal MWEs in Yiddish (Liebeskind & HaCohen-Kerner 2018)								
http://liebeskind-chaya.blogspot.co.il/p/downloads.html								
2018	YI	M	free	NO	NO	Can	NO	NO

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
870 English idioms: norming and statistical analysis (Bulkes & Tanner 2017)								
https://static-content.springer.com/esm/art								
2017	EN	M	free	NO	NO	Can	NO	Un
Czech MWEs (Nevěřilová 2018)								
https://live.european-language-grid.eu/catalogue/lcr/1200								
2017	CZ	M	free	NO	NO	Can	NO	NO
Czech Verbal MWEs (Bejček 2017)								
https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2603								
2017	CZ	S/A	free	YES	NO	Abs	NO	YES
NileULex (El-Beltagy 2016)								
https://live.european-language-grid.eu/catalogue/lcr/1374								
2017	AR	S/A	free	NO	NO	Can	POL	NO
ParaDi 2.0 dataset (Barančíková & Kettnerová 2017)								
https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2605								
2017	CZ	M	free	NO	NO	Can	SEMRELS	YES
Dictionary of Estonian Synonyms (Institute of the Estonian Language 2016)								
https://live.european-language-grid.eu/catalogue/lcr/13590								
2016	ET	M	Un	NO	NO	Can	SEMRELS; GLOSS; STYL	NO
LEX-MWE-PT: Word Combination in Portuguese (Antunes & Mendes n.d.)								
https://catalog.elra.info/en-us/repository/browse/ELRA-L0097/								
2016	PT	Un	paid	YES	NO	Can	PAR; GLOSS; SEMREL	YES

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
Lexical Resource of Hebrew Verb-Noun MWEs (Liebeskind & HaCohen-Kerner 2016)								
http://liebeskind-chaya.blogspot.co.il/p/downloads.html								
2016	HE	M	free	NO	NO	Abs	NO	NO
Multilingual Lexicon of Nominal Compound Compositionality (Ramisch 2016)								
https://pageperso.lis-lab.fr/carlos.ramisch/?page=downloads/compounds								
2016	EN, FR, PT	S/A	free	YES	NO	Can	COMPS; GLOSS	Un
MWEs Dataset for Indian Languages (Singh et al. 2016)								
https://www.cfilt.iitb.ac.in/								
2016	HI, MR	S/A	free	NO	NO	Can	GLOSS	YES
MWEs in Spanish Dialects (Bogantes et al. 2016)								
https://perso.limsi.fr/savary/resources.php								
2016	ES, dialects: ES-CO, ES-CR, ES-MEX, ES-PE	S/A	Un	YES	NO	Can	GLOSS	YES
SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal MWEs (Al-Badrashiny 2016)								
https://aclanthology.org/W16-5414/								
2016	AR	S/A	Un	NO	NO	Can+Abs	NO	YES

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
Dictionary of Estonian Phraseology (Õim 2000)								
https://eki.ee/keeleinfo/sonastikud/								
2015	ET	M	Un	NO	NO	Can	NO	NO
hrMWELex – Croatian lexicon of MWEs (Ljubešić et al. 2014)								
https://www.clarin.si/repository/xmlui/handle/11356/1177								
2015	HR	A	free	YES	YES	Can	NO	YES
slMWELex – Slovene lexicon of MWEs (Ljubešić et al. 2015)								
https://www.clarin.si/repository/xmlui/handle/11356/1179								
2015	SL	A	free	YES	YES	Can	NO	NO
srMWELex v0.5 – Serbian lexicon of MWEs (Ljubešić et al. 2015)								
https://www.clarin.si/repository/xmlui/handle/11356/1178								
2015	SR	A	free	YES	YES	Can	NO	YES
ConceptNet-el (Giouli 2023)								
https://inventory.clarin.gr/lcr/1600								
2014	EL	M	free	YES	NO	Can	GLOSS; SEMRELS; POL	YES
Dictionary of idioms for Georgian (Lobzhanidze 2019)								
https://idioms.iliauni.edu.ge/								
2014	KA, EL	S/A	free	NO	NO	Can	GLOSS	NO
Grammatical Dictionary of Polish MWEs (Czerepowicka & Savary 2015)								
https://sgjw.clarin-pl.eu/								
2014	PL	S/A	free	NO	NO	Can	NO	YES

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
Referentiebestand Belgisch-Nederlands (Dutch Language Institute 2016)								
https://live.european-language-grid.eu/catalogue/lcr/13397								
2014	NL	M	free	NO	NO	Can	NO	NO
Bulgarian MWE dictionary (Koeva et al. 2016)								
https://dcl.bas.bg/mwe-dictionary-data/								
2013	BG	Un	Un	NO	YES	Can	NO	Un
DiCE: Diccionario de Colocaciones del Español (Hedegaard & Troelsgård 2010)								
http://www.dicesp.com/paginas								
2012	ES	S/A	free	YES	NO	Can	LEXF; GLOSS	YES
Compound Noun Compositionality Dataset (Reddy et al. 2011)								
https://sivareddy.in/downloads.html								
2011	EN	M	free	NO	Un	Can	COMPOS; EXT-GLOSS	Un
Noun Compound Senses (NCS) dataset (Reddy et al. 2011)								
https://sivareddy.in/downloads/index.html								
2011	EN	S/A	free	YES	NO	Can	COMPOS	YES
Terminology database of expressions (ELRA 2010)								
https://live.european-language-grid.eu/catalogue/lcr/2853/overview/								
2010	EN, FR	M	paid	NO	NO	Un	SEMRELS	YES
Czech Dependency Bigrams from the Prague Dependency Treebank (Pecina 2008)								
https://live.european-language-grid.eu/catalogue/lcr/1086								
2008	CZ	M	free	NO	NO	Can	NO	YES
Russian Collocations Database (Khokhlova 2020)								

Table 2 continued

Name of the resource, reference, link								
Year	**Langs**	**Method**	**Access**	**Linked to corpus**	**Linked to other**	**MWEI**	**Meaning**	**Syntax**
https://collocations.spbu.ru/								
Un	RU	A	free	YES	YES	Can	EXTGLOSS	YES

Table 3: Distribution of different languages with their level of technical support (according to ELE report and Joshi et al. (2020)) and the number of MWE resources they are involved in. *Resources whose evaluation is not present in ELE report and is extracted from Joshi et al. (2020). **Pomak is not classified but there are very limited resources on it, thus we assume its support to be 'Weak or none.'

Language	Support	# resources
English	GOOD	15
Czech	FRAGMENTARY (HIGHER)	5
Spanish	MODERATE	5
Greek	FRAGMENTARY (HIGHER)	4
Portuguese	FRAGMENTARY (HIGHER)	4
Bulgarian	FRAGMENTARY (LOWER)	3
Croatian	FRAGMENTARY (LOWER)	3
Dutch	FRAGMENTARY (HIGHER)	3
French	MODERATE	3
Italian	FRAGMENTARY (HIGHER)	3
Polish	FRAGMENTARY (HIGHER)	3
Russian	MODERATE/4*	3
Slovene	FRAGMENTARY (LOWER)	3
Arabic	GOOD/5*	2
Chinese (ZH)	GOOD/5*	2
Estonian	FRAGMENTARY (HIGHER)	2
German	MODERATE	2
Japanese	GOOD/5*	2
Serbian	WEAK OR NONE	2
Swedish	FRAGMENTARY (HIGHER)	2
Basque	FRAGMENTARY (LOWER)	1
Finnish	FRAGMENTARY (HIGHER)	1
Georgian	FRAGMENTARY (HIGHER)/3*	1
Pomak	WEAK OR NONE**	1
Hebrew	FRAGMENTARY (HIGHER)/3*	1
Hindi	MODERATE/4*	1

Table 3 continued

Language	Support	# resources
Irish	WEAK OR NONE	1
Korean	MODERATE/4*	1
Lithuanian	FRAGMENTARY (HIGHER)	1
Marathi	FRAGMENTARY (LOWER)/2*	1
Persian	MODERATE/4*	1
Romanian	FRAGMENTARY (HIGHER)	1
Spanish, Argentina	UNKNOWN	1
Spanish, Chile	UNKNOWN	1
Spanish, Colombia	UNKNOWN	1
Spanish, Costa Rica	UNKNOWN	1
Spanish, Mexico	UNKNOWN	1
Spanish, Peru	UNKNOWN	1
Yiddish	WEAK OR NONE/1*	1