Survey on Lexical Resources Focused on Multiword Expressions for the Purposes of NLP

Verginica Barbu Mititelu
RACAI
Bucharest, Romania
vergi@racai.ro

Voula Giouli
Aristotle University of Thessaloniki
Greece
pgiouli@del.auth.gr

Gražina Korvel
Vilnius University
Vilnius, Lithuania
grazina.korvel@mif.vu.lt

Chaya Liebeskind
Jerusalem College of Technology
Jerusalem, Israel
liebchaya@gmail.com

Irina Lobzhanidze
Ilia State University
Tbilisi, Georgia
irina_lobzhanidze@iliauni.edu.ge

Rusudan Makhachashvili
B. Grinchenko Metropolitan Univ.
Kyiv, Ukraine
r.makhachashvili@kubg.edu.ua

Stella Markantonatou
ILSP and Archimedes Unit-RC ATHENA
Athens, Greece
marks@athenarc.gr

Alexandra Marković
Inst. for the Serbian Language SASA
Belgrade, Serbia
aleksandra.markovic@isj.sanu.ac.rs

Ivelina Stoyanova
Inst. for Bulgarian Language, BAS
Sofia, Bulgaria
iva@dcl.bas.bg

Abstract

Lexica of MWEs have always been a valuable resource for various NLP tasks. This paper presents the results of a comprehensive survey on multiword lexical resources that extends a previous one from 2016 to the present. We analyze a diverse set of lexica across multiple languages, reporting on aspects such as creation date, intended usage, languages covered and linguality type, content, acquisition method, accessibility, and linkage to other language resources. Our findings highlight trends in MWE lexicon development focusing on the representation level of languages. This survey aims to support future efforts in creating MWE lexica for NLP applications by identifying these gaps and opportunities.

1 Motivation

Multiword expressions (MWEs) pose a unique challenge in Natural Language Processing (NLP), primarily due to their semantic non-compositionality. This characteristic makes their automatic identification in text crucial for semantically driven downstream applications. Despite recent advances, including the advent of large (and small) language models, MWEs' inherent complexity and distributional properties continue to impede their effective processing. Lexical resources, that is, computational lexica dedicated to MWEs, are essential to address these challenges (Savary et al., 2019).

Our objective is to provide a comprehensive overview of the current landscape of MWE-related computational lexica that have been created for NLP purposes. The identification of relevant resources was meant to be as exhaustive as possible. Special emphasis was placed on the languages featured in the resources and their levels of representation in the NLP ecosystem. Thus, the survey aims to serve as a first step in highlighting the extent to which less-represented languages are included and supported in existing resources.

The paper is organized as follows. Section 2 presents previous surveys that focus on MWEs and outlines the new features offered by the current one. Section 3 discusses the sources and methodology we adopted to compile the list of resources with their relevant characteristics. The overview of the current landscape of MWE lexical resources is presented in Section 4, before concluding the paper by setting objectives for future work (Section 5).

2 Previous surveys

We are aware of four surveys heretofore focused on MWEs: Rosén et al. (2015) focused on the types of

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 41–57

MWEs that were more frequently annotated in treebanks at that time, namely named entities, phrasal verbs, and prepositional MWEs. Rosén et al. (2016) compared the way in which light verb constructions and verbal idioms were annotated in treebanks and proposed general guidelines for this. The survey by Mahajan et al. (2024) is focused on the methodologies and features required to implement MWE detection systems and is therefore of little relevance to our work.

Our survey builds on the one by 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, lexica (either dedicated to MWEs or including them alongside other lexical entries), and corpora such as treebanks available at that moment. The survey was based, on the one hand, on keyword querying of three language resource platforms: META-SHARE (Piperidis, 2012), ELRA[2] and SIGLEX-MWE[3]. On the other hand, the linguistic community was approached and asked to fill in a form about resources familiar to them.

General information about each resource was recorded, such as its name, a link to it, its type, contact information, the language(s) covered, its size, the maximum length of the contained MWEs, whether it includes non-contiguous 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 the end of 2024. Several resources published before 2016, either not included in the PARSEME survey or significantly updated after that, have also been added. Moreover, our survey expands the description of each lexical resource in terms of several criteria presented in Section 4.

3 Data collection

We aimed at a comprehensive collection of relevant data that would enable us to draw an accurate picture of the MWE resource landscape by cataloging MWE-related lexica and detailing their properties. To achieve this, we defined the criteria for resource inclusion, which focused on retaining only computational lexica, databases, and lists centered on MWEs while excluding corpora, terminological databases, and named entity lists, thus departing from the approach of Losnegaard et al. (2016), that considered both lexical resources and parsed corpora, i.e., treebanks, in their survey.

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

1. *European Language Grid (ELG)*[4] (Rehm, 2023). This is the largest platform where language technologies and language resources alike, developed by public or private bodies, are cataloged and stored to increase their visibility among potential users and developers and to facilitate access to them. The catalog 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 examined their description to decide upon their inclusion in the dataset.

2. *ACL Anthology*[5] 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 and was then checked by the authors.

3. *Europhras Conference Proceedings Repository*[6] provides lists of publications with relevant metadata. All publications after 2016 were checked. The resources retrieved overlapped with those from the ELG and ACL repositories.

4. *Phraseology and Multiword Expressions book series*[7] of Language Science Press was 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 provide a detailed description of the linguistic information and representation of MWEs.

[1] https://typo.uni-konstanz.de/parseme/
[2] https://www.elra.info/
[3] https://multiword.org/
[4] https://live.european-language-grid.eu/
[5] https://aclanthology.org/
[6] http://www.europhras.org/en/conferences
[7] https://langsci-press.org/catalog/series/pmwe

5. *Arxiv digital open access repository*[8] includes a wide range of scholarly articles in different areas. We have 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, mainly 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, p. 2302–03) 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.

4 MWE lexical resources: overview

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. The list records detailed information about each resource, such as publication date (or date of the last update), linguality (monolingual, bilingual, multilingual), resource type, acquisition method, licensing information, etc. These are extracted from the paper documenting the resource, from the resource website, or observed via manual resource inspection. The resources included in the survey are presented in Table 1 in the Appendix.

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 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 creation, update, or publication of the resources, focusing predominantly on lexical resources produced after 2016. Most identified resources are new; only three of them are enriched and updated. We also included several resources published before 2016 that were not included 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 the development of new MWE resources.[9] The distribution of resources by year of publication is plotted against relevant EU-funded initiatives for reference: META-NET Project[10], PARSEME COST Action[11], Horizon 2020 ELEXIS Project[12], UniDive COST Action[13].

4.2 Intended usage

The main inclusion criterion was 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

[8] https://arxiv.org/

[9] 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 have been included in the examined repositories at the moment of our investigation.

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

[11] https://typo.uni-konstanz.de/parseme/

[12] https://elex.is/

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

than the former: from the 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 (8 resources). Six resources are meant for 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. These lexica cover 37 languages (42 including varieties) in total. For comparison, the PARSEME survey included 14 bi- or multilingual resources (13% of the total resources 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), French (FR), German (DE), Modern Greek (EL) (3 lexica), Hebrew (HE), Irish (GA), Italian (IT), Lithuanian (LT), 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) (2), and Yiddish (YI). 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 10 lexica are bilingual, covering 12 languages (6 of which do not appear in monolingual resources) and 9 language pairs. Half of these are unidirectional from a source language to the target: Polish-English (PL-EN), English-French (EN-FR), English-Italian (EN-IT), English-Persian (EN-FA), Georgian-Modern Greek (KA-EL), Croatian-English (HR-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. Finally, one resource involves two Indian language varieties, namely Hindi (HI) and Marathi (MR) – yet they are not aligned as translation dictionaries. Finally, 5 resources are multilingual, covering 10 languages in all (3 out 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 5 languages, namely English (EN), German (DE), Italian (IT), Portuguese (PT), and Russian (RU), while another resource covers 4 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. In 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, 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 linguistic descriptions of MWE-related phenomena are 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%), 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 of the resources were reportedly developed either semi-automatically (26 resources or 39.9%) or fully automatically (8 resources or 12.1%). In 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 innovation 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 (3 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 even (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

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.[14]

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

[14] 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 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).

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 an existing correlation between language typology and the level of language resourcefulness. In short, they divide languages into six classes (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); 4 – Moderate; 5 – Good).[15] 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 (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 2 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 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 that, while the community continues to develop MWE lexical resources for languages with good and moderate support, in recent years (since 2016), the focus has predominantly shifted toward compiling MWE lexica for lower-resourced languages (with fragmentary, weak, or no support). Furthermore, there has been extensive work on MWE lexica for non-European languages and language varieties, particularly 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 most linked to corpora or other data sources. In 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.

Lexica linked to corpora are predominantly derived automatically (27.3%) or semiautomatically (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

[15] There are some discrepancies in the alignment for two languages, namely for 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/.

or no linguistic description and no confirmed MWE status, which are not, as mentioned a few times, included in the present survey.

5 Conclusions and outlook for future research

We set out for this survey by examining several features of existing MWE resources. So, we described the macro-properties of computational lexica, such as linguality, availability, acquisition method, and linkage to external general lexica. In this section, we summarize and discuss our findings.

Regarding linguality, most lexica are either 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, Persian, Hebrew, and Korean. 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 NLP-involved domains.

Most resources are MWE-dedicated, but some present both one-word and multi-word entries. MWE-dedicated resources are generally independent and not linked to general lexica. Resources tend to address the general language with few exceptions, such as lexica for specific purposes, i.e., expressions denoting sentiments. Again, a phenomenon of neglecting within-language diversity is observed as these (predominantly colloquial) language aspects have not been 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. However, the description of the contents of the resources often lacks the detail and clarity required to understand precisely what type of information the resources offer.

Most of the resources were developed (semi-)automatically. However, the literature does not provide benchmarks or diagnostics for measuring the quality of resources, whether created automatically or manually.

The size of the resources varies, but generally, resources are not big; this might indicate the effort required to develop such resources. We chose not to include any information on the size of the resources since it is not uniformly documented or the size information is entirely missing.

Our survey has highlighted the role of EU-funded projects related to lexical resources, such as the COST actions PARSEME and UniDive, and Horizon-funded project ELEXIS. These initiatives, as well as the European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))[16], have contributed significantly to the development of resources (see Figure 1). Related to this is the observation that more MWE resources have been developed recently for less-resourced languages rather than well-resourced ones. Although this might be due, among others, to the fact that well-resourced languages already possess MWE resources, one should consider that EU initiatives such as the ones listed above provide special encouragement for studying less-resourced languages and language varieties, in line with the EU priority to preserve multilinguality in Europe.

Our recommendations regarding the macroscopic properties of lexica are:
– Document the design and the contents of the resources thoroughly, clearly, and concisely.
– Make the resource freely available, at least for research purposes.
– Make the resource accessible through stable and friendly repositories.
– Ensure resource maintenance over time.
– Cover special usages of language, such as offensive speech.

In our future research, we will further explore the types of (linguistic) information about MWEs provided by these resources and the way in which it is described. We will further try to identify the best encoding practices.

6 Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). Also, part of this research was supported by Aristotle University of Thessaloniki (Grant ELKE-AUTH-

[16] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018IP0332

13337) and the Ministry of Science, Republic of Serbia #GRANT 451-03-66/2024-03/200174.

References

Mohamed Al-Badrashiny. 2016. SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal Multiword Expressions Tokens Paradigm and their Morphosyntactic Features. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 113–122.

Valentin Anders. 2022a. Chilenismos (deChile).

Valentin Anders. 2022b. Expressions (deChile).

Sandra Antunes and Amália Mendes. 2013. MWE in Portuguese: Proposal for a Typology for Annotation in Running Text. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 87–92, Atlanta, Georgia, USA. Association for Computational Linguistics.

Petra Barančíková and Václava Kettnerová. 2017. ParaDi: Dictionary of Paraphrases of Czech Complex Predicates with Light Verbs. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Eduard Bejček. 2017. Czech Verbal MWEs. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL).

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

Goranka Blagus Bartolec, Gorana Duplančić Rogošić, and Antonia Ordulj. 2024. INIKOL - Collocational Database for Learning Croatian as a Foreign Language. In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 8–12, Gothenburg, Sweden. Association for Computational Linguistics.

Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, and Agata Savary. 2016. Towards Lexical Encoding of Multi-Word Expressions in Spanish Dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2255–2261, Portorož, Slovenia. European Language Resources Association (ELRA).

Nyssa Z. Bulkes and Darren Tanner. 2017. "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, 49(2):772–783.

Monika Czerepowicka and Agata Savary. 2015. SEJF -a Grammatical Lexicon of Polish Multi-Word Expressions. In *Proceedings of the Language Technology Conference 2015 (LTC 2015)*, page 5, Poznań, Poland.

Dutch Language Institute. 2016. Referentiebestand Belgisch-Nederlands. European Language Grid.

Samhaa R. El-Beltagy. 2016. NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2900–2905, Portorož, Slovenia. European Language Resources Association (ELRA).

ELRA. 2010. Terminology database of expressions.

ELRA. 2019. English-Persian database of idioms and expressions.

Ivana Filipović Petrović, Miguel López Otal, and Slobodan Beliga. 2024. Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4106–4112, Torino, Italia. ELRA and ICCL.

Beatriz Fisas. 2020. CollFrEn: Rich Bilingual English–French Collocation Resource. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 1–12.

Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. 2023. Digital language equality: Definition, metric, dashboard. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, pages 39–73. Springer International Publishing, Cham.

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

Jette Hedegaard and Thomas Troelsgård. 2010. Dice in the Web: an Online Spanish Collocation Dictionary. In *Elexicography in the 21st century: new challenges, new applications. Proceedings of ELEX2009, Louvain-la-Neuve, 22-24 October 2009*, pages 369–374. Cahiers Du Cental 7. Louvain-la-Neuve, Presses Universitaires de Louvain.

Ferdy Hubers, Catia Cucchiarini, and Helmer Strik. 2020. Dedicated Language Resources for Interdisciplinary Research on Multiword Expressions: Best Thing since Sliced Bread. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4418–4425, Marseille, France. European Language Resources Association.

Institute of the Estonian Language. 2016. The Dictionary of Estonian Synonyms. European Language Grid.

Uxoa Iñurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2018. Konbitzul: an MWE-specific database for Spanish-Basque. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Charles Jochim. 2018. SLIDE - a Sentiment Lexicon of Common Idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kyoko Kanzaki. 2019. Towards linking synonymous expressions of compound verbs to Japanese WordNet. In *Proceedings of the 10th Global Wordnet Conference*, pages 185–190.

Maria Khokhlova. 2020. Collocations in Russian Lexicography and Russian Collocations Database. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3198–3206. European Language Resources Association.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, and 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*, pages 86–95.

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

Cvetana Krstev, Ivan Obradović, Ranka Stanković, and Duško Vitas. 2013. An Approach to Efficient Processing of Multi-word Units. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz, editors, *Computational Linguistics*, volume 458, pages 109–129. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Studies in Computational Intelligence.

Murathan Kurfalı, Robert Östling, Johan Sjons, and Mats Wirén. 2020. A Multi-word Expression Dataset for Swedish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4402–4409, Marseille, France. European Language Resources Association.

Svetlozara Leseva, Verginica Barbu Mititelu, and Ivelina Stoyanova. 2020. It Takes Two to Tango – Towards a Multilingual MWE Resource. In *Proceedings of the Fourth International Conference on Computational Linguistics in Bulgaria (CLIB 2020)*, pages 101–111, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *arXiv preprint*. ArXiv:2308.13961 [cs].

Chaya Liebeskind and Yaakov HaCohen-Kerner. 2016. A Lexical Resource of Hebrew Verb-Noun Multi-Word Expressions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 522–527, Portorož, Slovenia. European Language Resources Association (ELRA).

Chaya Liebeskind and Yaakov HaCohen-Kerner. 2018. Verbal Multi-Word Expressions in Yiddish. In *Natural Language Processing and Information Systems*, pages 205–216, Cham. Springer International Publishing.

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

Nikola Ljubešić, Kaja Dobrovoljc, and Darja Fišer. 2015. *MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica*, 39(3).

Irina Lobzhanidze. 2019. Computational Model of the Modern Georgian Language and Search Patterns for an Online Dictionary of Idioms. In A. Silva, S. Staton, P. Sutton, and C. Umbach, editors, *Language, Logic, and Computation*, volume 11456 of *Lecture Notes in Computer Science*, pages 187–208. Springer.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia. European Language Resources Association (ELRA).

Ujwala P. Mahajan, Ajay S. Patil, and Nita V. Patil. 2024. A survey of tools and techniques for multiword expression detection. *International Journal of Computer Applications*, 186(32):11–18.

Stella Markantonatou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 130–134. Association for Computational Linguistics.

Stella Markantonatou, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chirila, Dimitrios Karamatskos, Nicolaos Valeontisa, and George Pavlidis. 2024. Description of Pomak within IDION: Challenges in the representation of verb multiword expressions. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Language Science Press.

Francesca Masini, M. Silvia Micheli, Andrea Zaninello, Sara Castagnoli, and Malvina Nissim. 2020. MWE_combinet_release_1.0. Associazione Italiana di Linguistica Computazionale.

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

Marek Maziarz, Łukasz Grabowski, Tadeusz Piotrowski, Ewa Rudnicka, and Maciej Piasecki. 2023. Lexicalisation of Polish and English word combinations: an empirical study. *Poznan Studies in Contemporary Linguistics*, 59(2):381–406. Publisher: De Gruyter Mouton.

Diego Moussallem. 2018. LIdioms: A Multilingual Linked Idioms Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zuzana Nevěřilová. 2018. Discovering Continuous Multi-word Expressions in Czech. *Computación y Sistemas*, 22(3).

Jan Odijk and Martin Kroon. 2024. A Canonical Form for Flexible Multiword Expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 91–101, Torino, Italia. ELRA and ICCL.

Petya Osenova and Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Language Science Press.

Irene Pagliai. 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.

Pavel Pecina. 2008. Gold Standard Reference Data for Multiword Expression Extraction: Czech Dependency Bigrams from the Prague Dependency Treebank. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL).

Stelios Piperidis. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 36–42, Istanbul, Turkey. European Language Resources Association (ELRA).

Dmitry Puzyrev, Artem Shelmanov, Alexander Panchenko, and Ekaterina Artemova. 2019. A Dataset for Noun Compositionality Detection for a Slavic Language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Carlos Ramisch. 2016. How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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

Frankie Robertson. 2020. Filling the ___-s in Finnish MWE lexicons. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 13–21. Association for Computational Linguistics.

Barrios Rodriguez and Maria Auxiliadora. 2019. A Spanish E-dictionary of Collocations. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 160–167.

Adolfo Enrique Rodríguez et al. 2022. Lunfardo Dictionary. European Language Grid.

Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. 2016. MWEs in treebanks: From survey to guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2323–2330, Portorož, Slovenia. European Language Resources Association (ELRA).

V. Rosén, G. S. Losnegaard, K. De Smedt, E. Bejček, A. Savary, A. Przepiórkowski, P. Osenova, and V. Barbu Mititelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 179—-193, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary and Silvio Ricardo Cordeiro. 2017. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 64–72, Prague, Czech Republic.

Fran Ramovš Institute of the Slovenian Language ZRC Sazu. 2023. Terminological multiword expressions lexicon. *https://slovenscina.eu/*.

Sabine Schulte im Walde. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*. Language Science Press.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016. Multiword Expressions Dataset for Indian Languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335, Portorož, Slovenia. European Language Resources Association (ELRA).

Amalia Todirascu. 2019. PolylexFLE : une base de données d'expressions polylexicales pour le FLE (PolylexFLE : a database of multiword expressions for French L2 language learning). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume I : Articles longs*, pages 143–156. ATALA.

Elena Volodina, David Alfter, and Therese Lindström Tiedemann. 2024. Profiles for Swedish as a Second Language: Lexis, Grammar, Morphology. In *Proceedings of the Huminfra Conference (HiC 2024), Gothenburg, 10–11 January 2024*, pages 10–19.

Pavel Vondřička. 2019. Design of a Multiword Expressions Database. *The Prague Bulletin of Mathematical Linguistics*, 112(1):83–101.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2019. Ilfhocail: A Lexicon of Irish MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 162–168, Florence, Italy. Association for Computational Linguistics.

Rodrigo Wilkens, Leonardo Zilio, Silvio Ricardo Cordeiro, Felipe Paula, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2017. LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.

Ziheng Zeng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264, Singapore. Association for Computational Linguistics.

Asta Õim. 2000. *Fraseoloogiasõnaraamat*, 2-ne, täiendatud ja parandatud trükk edition. Eesti keele sihtasutus, Tallinn.

Appendix

Table 1: List of resources included in the survey with basic reference information. Method: M – manual processing; S/A – semi-automatic; A – automatic.

Lexicon	Link	Reference	Langs	Method	Access
LEMUR	→	(Vondřička, 2019)	CZ	S/A	unclear
NileULex	→	(El-Beltagy, 2016)	AR	S/A	free
LEX-MWE-PT: Word Combination in Portuguese	→	(Antunes and Mendes, 2013)	PT	other-unclear	paid
Lexicalisation of Polish and English word combinations	→	(Maziarz et al., 2023)	PL, EN	M	free
The Database of Lithuanian MWEs	→	(Bielinskienė et al., 2022)	LT	S/A	free
srMWELex v0.5 – Serbian lexicon of MWEs	→	(Ljubešić et al., 2015)	SR	A	free
hrMWELex – Croatian lexicon of MWEs	→	(Ljubešić et al., 2014)	HR	A	free
slMWELex – Slovene lexicon of MWEs	→	(Ljubešić et al., 2015)	SL	A	free
Srp_DELAC	→	(Krstev et al., 2013)	SR	M	academic
Expressions (deChile)	→	(Anders, 2022b)	ES	M	free
Czech MWEs	→	(Nevěřilová, 2018)	CZ	M	free
Dictionary of Estonian Phraseology	→	(Õim, 2000)	ET	M	unclear
Terminological MWE lexicon	→	(Sazu, 2023)	SL	M	free
Terminology database of expressions	→	(ELRA, 2010)	EN, FR	M	paid
Idioms of Chile [Chilenismos]	→	(Anders, 2022a)	ES-CL	M	free
Lunfardo Dictionary	→	(Rodríguez et al., 2022)	ES-AR	M	free
Dictionary of Estonian Synonyms	→	(Institute of the Estonian Language, 2016)	ET	M	unclear
ilFhocail	→	(Walsh et al., 2019)	GA	M	unclear
Referentiebestand Belgisch-Nederlands	→	(Dutch Language Institute, 2016)	NL	M	free
Czech Dependency Bigrams from the Prague Dependency Treebank	→	(Pecina, 2008)	CZ	M	free
Konbitzul	→	(Iñurrieta et al., 2018)	EU, ES	M	free
English-Persian database of idioms and expressions	→	(ELRA, 2019)	EN, FA	S/A	paid
ParaDi 2.0 dataset	→	(Barančíková and Kettnerová, 2017)	CZ	M	free
MWE lexicon extracted from the Gigafida 2.1 corpus	→	(Krek et al., 2021)	SL	S/A	unclear
Czech Verbal MWEs	→	(Bejček, 2017)	CZ	S/A	free
Bulgarian MWE dictionary	→	(Koeva et al., 2016)	BG	unclear	unclear
ConceptNet-el	→	(Giouli, 2023)	EL	M	free
CollFrEn: Rich Bilingual English–French Collocation Resource	→	(Fisas, 2020)	EN, FR	A	free
FinnMWE: a lexicon of Finnish MWEs	→	(Robertson, 2020)	FI	A	free
Russian Collocations Database	→	(Khokhlova, 2020)	RU	A	free
Diretes (DIccionario RETicular de ESpañol)	→	(Rodriguez and Auxiliadora, 2019)	ES	unclear	unclear
IDION: A database for Modern Greek MWEs		(Markantonatou, 2019)	EL	unclear	free
PolylexFLE	→	(Todirascu, 2019)	FR	unclear	unclear

Japanese compound verb lexicon	→	(Kanzaki, 2019)	JA, EN, ZH, KO	other-unclear	free
Sentiment Lexicon of IDiomatic Expressions (SLIDE)	→	(Jochim, 2018)	EN	S/A	free
LIDIOMS: A Multilingual Linked Idioms Data Set	→	(Moussallem, 2018)	EN, DE, IT, PT, RU	S/A	free
LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds	→	(Wilkens et al., 2017)	PT	S/A	unclear
SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal MWEs	→	(Al-Badrashiny, 2016)	AR	S/A	unclear
Multilingual Lexicon of Nominal Compound Compositionality	→	(Ramisch, 2016)	EN, FR, PT	S/A	free
Lexical Resource of Hebrew Verb-Noun MWEs	→	(Liebeskind and HaCohen-Kerner, 2016)	HE	M	free
MWEs in Spanish Dialects	→	(Bogantes et al., 2016)	ES, dialects: ES-CO, ES-CR, ES-MEX, ES-PE	S/A	unclear
MWEs Dataset for Indian Languages	→	(Singh et al., 2016)	HI, MR	S/A	free
Noun Compound Senses (NCS) dataset	→	(Reddy et al., 2011)	EN	S/A	free
MWE Dataset for Swedish	→	(Kurfalı et al., 2020)	SV	S/A	free
Noun Compound Dataset for Russian	→	(Puzyrev et al., 2019)	RU	S/A	free
Diccionario de Colocaciones del Español (DiCE)	→	(Hedegaard and Troelsgård, 2010)	ES	S/A	free
Polish verbal MWEs	→	(Savary and Cordeiro, 2017)	PL	A	free
Dutch idiomatic expressions		(Hubers et al., 2020)	NL	A	free
MWE_combinet_release_1.0	→	(Masini et al., 2020)	IT	S/A	free
Grammatical Dictionary of Polish MWEs	→	(Czerepowicka and Savary, 2015)	PL	S/A	free
Dictionary of idioms for Georgian	→	(Lobzhanidze, 2019)	KA, EL	S/A	free
IDION POMAK	→	(Markantonatou et al., 2024)	POMAK	M	free
DUCAME	→	(Odijk and Kroon, 2024)	NL	unclear	unclear
MWE dictionary for Bulgarian and Romanian		(Leseva et al., 2020)	BG, RO	S/A	free
Feature-NN	→	(Schulte im Walde, 2024)	DE	S/A	free
Compound Noun Compositionality Dataset	→	(Reddy et al., 2011)	EN	M	free
MWE-CEFR Profiles	→	(Volodina et al., 2024)	SV	S/A	free
Bulgarian Integrated Lexicon	TBA	(Osenova and Simov, 2024)	BG		
MWEs in FrameNet-EL	TBA		EL		
Verbal MWEs in Yiddish	→	(Liebeskind and HaCohen-Kerner, 2018)	YI	M	free
IdiomKB	→	(Li et al., 2023)	EN, ZH, JA	S/A	free
870 English idioms: norming and statistical analysis	→	(Bulkes and Tanner, 2017)	EN	M	free
Collocational Database for Learning Croatian as a Foreign Language		(Blagus Bartolec et al., 2024)	HR, EN	other-unclear	free

Normed lexicon of English and Italian idioms	→	(Pagliai, 2023)	EN, IT	unclear	free for specific uses
Croatian dictionary of idioms	→	(Filipović Petrović et al., 2024)	HR	S/A	free
IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions	→	(Zeng et al., 2023)	EN	S/A	free

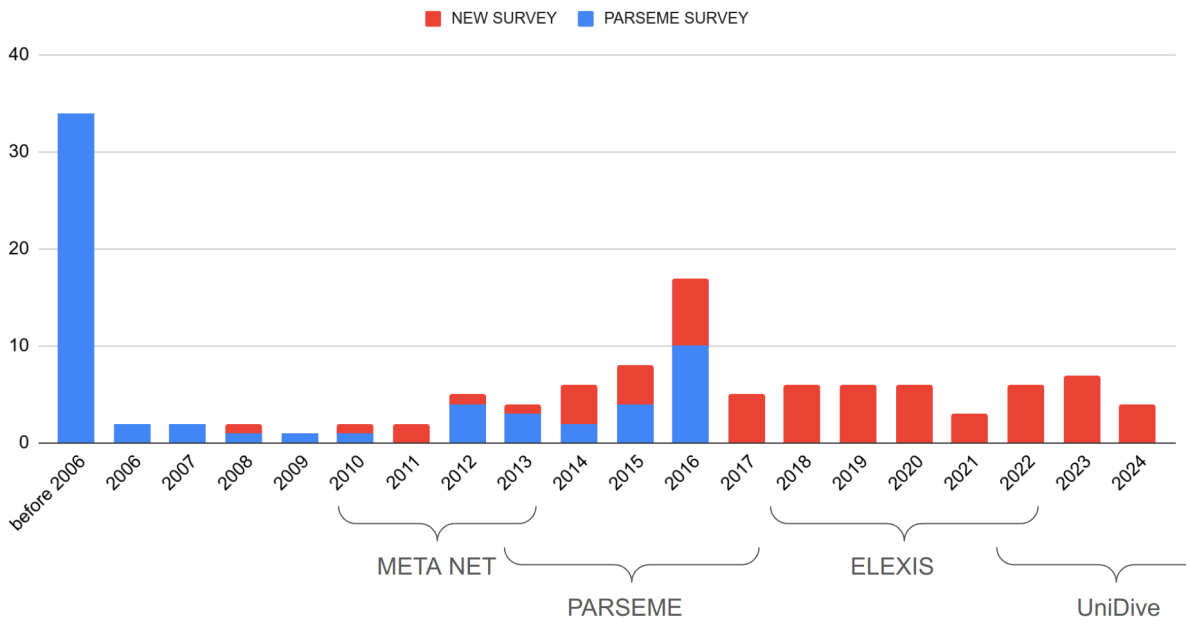

Figure 1: Distribution of resources by year of publication

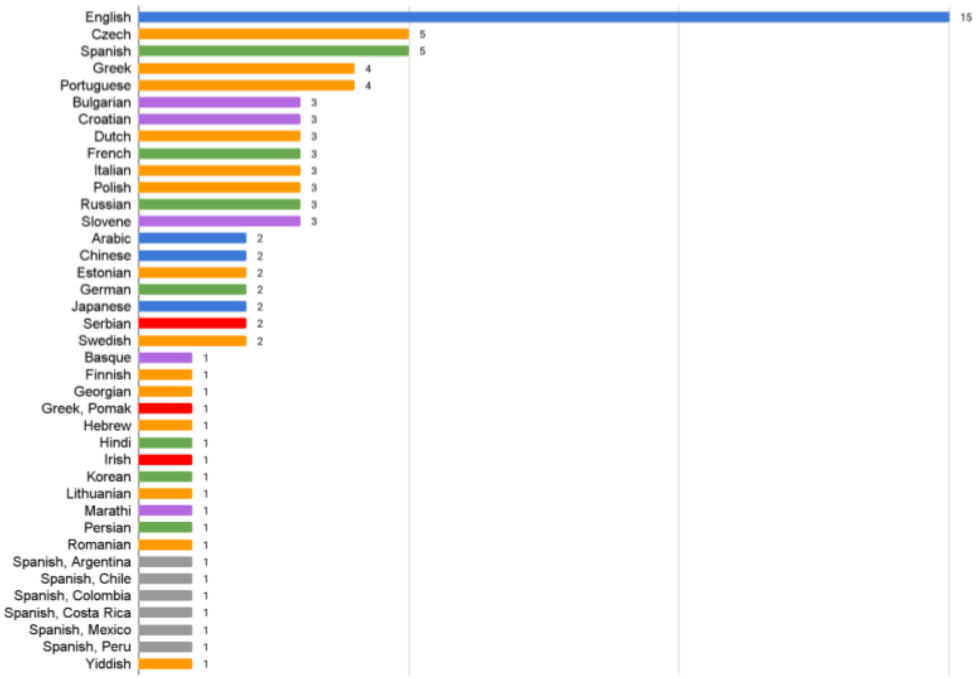

Figure 2: Distribution of different languages and the number of MWE resources they are involved in. The color shows the level of technical support: blue – Good; green – Moderate; orange – Fragmentary (higher); purple – Fragmentary (lower); red – Weak or none; gray – Unknown.

Table 2: 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 (ELE report)**	**# resources**	**Language**	**Support (ELE report)**	**# resources**
English	GOOD	15	Basque	FRAGM (LOWER)	1
Czech	FRAGM (HIGHER)	5	Finnish	FRAGM (HIGHER)	1
Spanish	MODERATE	5	Georgian	FRAGM (HIGHER)/3*	1
Greek	FRAGM (HIGHER)	4	Pomak	WEAK OR NONE**	1
Portuguese	FRAGM (HIGHER)	4	Hebrew	FRAGM (HIGHER)/3*	1
Bulgarian	FRAGM (LOWER)	3	Hindi	MODERATE/4*	1
Croatian	FRAGM (LOWER)	3	Irish	WEAK OR NONE	1
Dutch	FRAGM (HIGHER)	3	Korean	MODERATE/4*	1
French	MODERATE	3	Lithuanian	FRAGM (HIGHER)	1
Italian	FRAGM (HIGHER)	3	Marathi	FRAGM (LOWER)/2*	1
Polish	FRAGM (HIGHER)	3	Persian	MODERATE/4*	1
Russian	MODERATE/4*	3	Romanian	FRAGM (HIGHER)	1
Slovene	FRAGM (LOWER)	3	Spanish, Argentina	UNKNOWN	1
Arabic	GOOD/5*	2	Spanish, Chile	UNKNOWN	1
Chinese (ZH)	GOOD/5*	2	Spanish, Colombia	UNKNOWN	1
Estonian	FRAGM (HIGHER)	2	Spanish, Costa Rica	UNKNOWN	1
German	MODERATE	2	Spanish, Mexico	UNKNOWN	1
Japanese	GOOD/5*	2	Spanish, Peru	UNKNOWN	1
Serbian	WEAK OR NONE	2	Yiddish	WEAK OR NONE/1*	1
Swedish	FRAGM (HIGHER)	2			

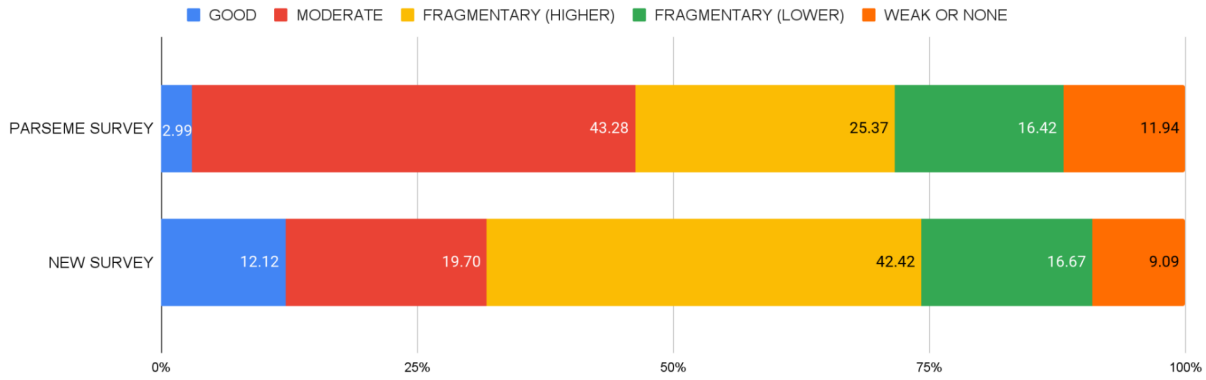

Figure 3: Distribution of resources according to level of technical support

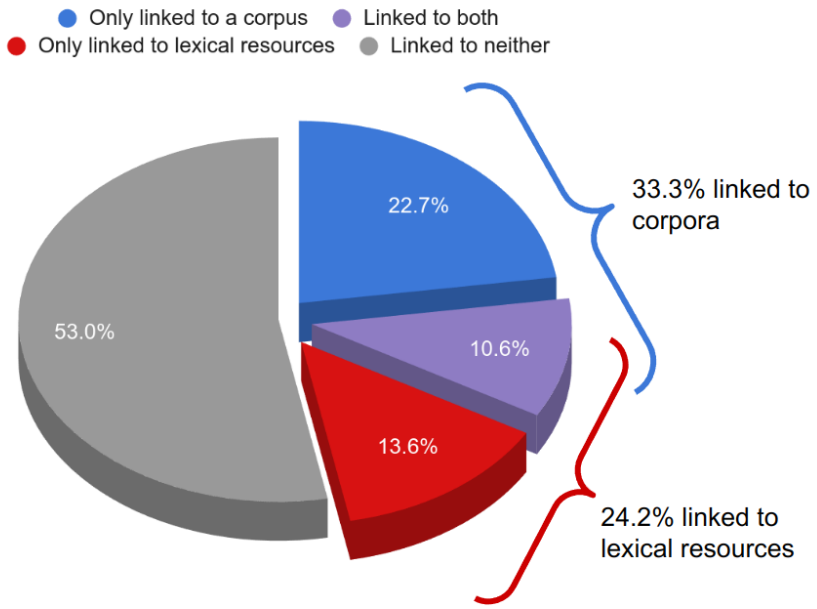

Figure 4: Distribution of resources according to their links to corpora and/or other lexical resources