Chapter 3

Using Multiword Expression lexicons in Natural Language Processing tasks: A survey

Voula Giouli[a], Verginica Barbu Mititelu[b], Raquel Amaro[c], Svetla Koeva[d], Gražina Korvel[e], Irina Lobzhanidze[f] & Giedre Valunaite Oleskeviciene[g]

[a]Aristotle University of Thessaloniki [b]Romanian Academy Research Institute for Artificial Intelligence [c]NOVA University Lisbon [d]Bulgarian Academy of Sciences Institute for Bulgarian Language [e]Vilnius University [f]Ilia State Unviersity [g]Mykolas Romeris University

Multiword expressions remain a persistent challenge for Natural Language Processing tools and applications across a wide range of tasks. This chapter presents and discusses the results of a twofold survey: it combines a comprehensive review of recent literature and a questionnaire for researchers and practitioners in the field. Both show which tasks/applications multiword expression lexicons are used for, what types of information from such resources are relevant for the respective tasks, what types of information could be worth including therein, and to what extent the use of the lexical resource at hand had an impact on the overall performance of the task or of the tool. The survey reveals that MWE lexicons are already used in core NLP tasks like parsing, machine translation, LLM evaluation and others. But in spite of that, richer and standardised annotation schemes and stronger links between lexicons and corpora are still needed to support their effectiveness.

1 Introduction

After more than two decades of research, Multiword Expressions (MWEs) remain a persistent challenge for Natural Language Processing (NLP) tools and



Voula Giouli, Verginica Barbu Mititelu, Raquel Amaro, Svetla Koeva, Gražina Korvel, Irina Lobzhanidze & Giedre Valunaite Oleskeviciene. 2026. Using Multiword Expression lexicons in Natural Language Processing tasks: A survey. In Verginica Barbu Mititelu & Voula Giouli (eds.), *Multiword expressions in Natural Language Processing: Current trends and challenges*, 99–153. Berlin: Language Science Press. DOI: 10.5281/zenodo.21277673

applications across a wide range of tasks. MWEs such as idioms, collocations, and phrasal verbs – among others – often do not adhere to the general rules of the language. Challenges are particularly evident in downstream tasks such as syntactic parsing, where MWEs disrupt grammatical structures, and semantic parsing, where their non-compositional meanings are hard to capture. Moreover, in certain applications such as machine translation or sentiment analysis, MWEs frequently lead to inaccurate results. To this end, MWE processing is of paramount importance, and it often relies on elaborate language resources, such as annotated corpora and computational lexicons (Constant et al. 2017).

This chapter presents and discusses the results of a twofold survey on the use of MWE lexicons in NLP tasks and applications. More precisely, our goal is to identify what types of lexicons and what kind of information contained in them have been employed in NLP tasks. In particular, we show: (i) which tasks/applications MWE lexicons are used for (Sections 5.3.1 and 6.2.1); (ii) what types of information from the MWE lexicons are relevant for the respective tasks (Sections 5.3 and 6.2.2); (iii) what types of information MWE lexicons lack and could be worth including (Section 6.2.3); and (iv) to what extent has the inclusion of the lexical resource at hand impacted the overall performance of the task or of the tool (Section 5.5). In other words, we want to quantify the contribution of MWE lexicons to the tasks and applications at hand, especially in the era of Machine Learning (ML) and Large Language Models (LLMs).

The contributions of this chapter are the following: (i) to the best of our knowledge, we deliver the first critical assessment of the impact of MWE lexicons on NLP tools and applications; (ii) we provide a thorough literature review on MWE lexicons and their contribution to NLP tasks and applications; (iii) complementarily, we present the results of a questionnaire, devised to gather updated information on how MWE lexicons are used in research and language technologies; and (iv) we identify characteristics and features that might be useful for future inclusion in MWE lexicons, as well as relevant knowledge that can be used to discover and inform synergies between efforts to build these resources and the development of NLP.

The remainder of the chapter is structured as follows: in Section 2, we present the scope and aims of the survey; Section 3 overviews relevant literature while Section 4 elaborates on the methodology adopted; Section 5 presents the findings of the critical review of prior work concerning the use of MWE-dedicated or MWE-aware lexicons across a range of NLP tasks; Section 6 reports on the user survey, considering dissemination, respondent profiles, and the data and results obtained; the findings of these surveys and their relevance for further developments of MWE lexicons are discussed in Section 7. Finally, Section 8 offers concluding remarks.

2 Scope and aims of the survey

Our study aims to explore the use of MWE-related lexical resources in NLP tasks. The intersection of the two, lexical resources and NLP tasks, is the focus of interest. Consequently, MWE lexical resources that are not used in NLP tasks fall outside the scope of the current research. In this section, we elaborate on the types of resources that are central to the survey and clarify the intended NLP tasks.

We are primarily interested in computational lexicons, that is, structured resources – often stored in a database – that catalog lexical entries (single- and/or multiword ones) with their linguistic properties or features at various levels of linguistic analysis (lexis, morphology, syntax, semantics, and pragmatics). The term computational lexicon refers here to a lexical resource in electronic form in which information about words and phrases is encoded. Computational lexicons appear in a variety of forms, ranging from simple text documents to more elaborate lexical databases, and are used to facilitate various NLP tasks. We are interested in MWE-dedicated and MWE-aware lexical resources (Losnegaard et al. 2016). While the former treat only MWEs as lexical entries and encode their idiosyncrasies, the latter are more inclusive in scope; they not only account for MWEs explicitly but also focus on lexemes of a language and integrate MWEs alongside single words within the same lexical framework. These lexicons provide representations of both single- and multiword entries, treating the latter as an integral part of the broader lexicon.

In the past, computational lexicons of all sorts were primarily designed as static resources tailored to specific applications. Their creation, maintenance, and enrichment relied heavily on manual or semi-automatic input from expert lexicographers to define and categorise MWEs. These lexicons typically encode linguistic information about lemma or canonical form, part-of-speech tags, syntactic structures, and even semantic features. However, with the advent of word embeddings and neural models, a shift has occurred towards data-driven approaches. Word embeddings, such as those created by word2vec or BERT, represent words and MWEs as dense vectors in a continuous space. This shift has opened new avenues for capturing the nuances of MWEs, as their meanings can be inferred through contextualized representations learned from large datasets. These models often show improvements in capturing the non-compositional nature of MWEs.

Recent developments also include the use of MWE-dedicated resources broadly referred to as “compositionality rating datasets”, which quantify how compositional or idiomatic a given MWE is. Compositionality ratings refer to

human evaluations of the degree to which the meaning of a compound word or phrase can be inferred from the meanings of its components (Reddy et al. 2011). There are two main approaches: (a) rating the compositionality of the entire compound, and (b) rating the degree of compositionality of the entire MWE in relation to its constituents. Such ratings are performed by expert annotators or by non-experts (native speakers of the language at hand via crowdsourcing, or a combination of both Reddy et al. 2011, Schulte im Walde et al. 2016, Schulte im Walde 2024). While initial attempts aimed at obtaining binary judgments, ratings are now typically collected using a scale: on one end of the continuum, compounds are judged as "completely opaque", while on the other end MWEs are rated as "completely transparent", with various degrees of partial transparency in between. These datasets can be used to train models that more effectively predict whether a MWE should be treated as a single lexical unit, or whether its meaning can be derived from its parts.

The focus of this study is placed on computational lexical resources of all sorts: computational lexicons, compositionality datasets, and resources aimed at contextualised representation of MWEs. We are equally interested in lexical resources that are manually constructed or obtained (semi-)automatically, provided they are used in NLP tasks. We also consider any new type of resources that reflect the trends of the new methodologies such as LLMs.

We delve into two types of NLP tasks. On the one hand, we are interested in the task of MWE identification; in this sense, the focus is on MWE lexical resources used within the task of MWE identification (Constant et al. 2017). On the other hand, the focus is on various NLP tasks that range from downstream ones (e.g., dependency parsing, semantic analysis) to applications such as information extraction, sentiment analysis, hate speech detection, machine translation, etc. Other uses of MWE lexicons by human users, such as in traditional language teaching and learning, while interesting, fall outside the scope of the current research, while usage of MWE lexicons for computer-assisted language learning is retained.

Moreover, in the era of Machine Learning, the emphasis is placed on LLMs optimized to represent spans of text and embeddings of all sorts of MWEs, as well as on all kinds of tasks using LLMs related to MWEs and their respective datasets.

3 Related work

Over the past ten years, surveys on MWEs have primarily focused on two main areas: (i) how MWEs are represented in various types of language resources, par-

ticularly with respect to language coverage, language diversity, and MWE metadata descriptions; and (ii) how MWEs are handled in NLP applications, either by analyzing their interaction with downstream tasks or by focusing on tools and techniques specifically designed to process MWEs.

In this section, we discuss relevant work aligned with these two areas and highlight how the present survey offers a distinct contribution, providing new insights into MWEs from a different perspective.

3.1 MWEs in language resources

Rosen et al. (2015) present a survey on the annotation of MWEs in 17 treebanks in 15 languages with the aim of contributing to the development of common principles and guidelines for the annotation of MWEs in syntactic resources. The survey was conducted by inviting experts familiar with each treebank to complete an online questionnaire. Respondents were asked to describe the annotation of different MWE types – such as nominal, phrasal, prepositional, adjectival, idiomatic, and other expressions – within the context of their specific language and theoretical framework, including whether the treebank followed a dependency or constituency model, and whether it was based on Head-driven Phrase Structure Grammar or Lexical Functional Grammar.

For each MWE type, participants reported whether the construction occurred in the language, whether it was annotated in the treebank, and whether it was analyzed compositionally, and were asked to provide example sentences along with analysis descriptions and search expressions.

The survey results were analyzed with a comparative focus on key categories of MWEs. Prepositional MWEs were found to be annotated in all surveyed resources, although the strategies varied considerably; these expressions were typically represented with internal syntactic structure rather than as words with spaces. Verb-particle constructions, such as phrasal verbs, were annotated in similar ways across treebanks. However, the main challenge associated with this category was not annotation strategy, but the difficulty in consistently identifying such constructions. Multiword named entities displayed the greatest variation in annotation, ranging from simple, fixed expressions to complex structures that sometimes embedded other named entities.

The study concluded that, while some MWEs are clearly language-specific, others – particularly named entities – are subject to more diverse and inconsistent annotation practices. It also highlighted a general lack of standardised annotation guidelines and metadata documentation across treebanks. These findings

underscore the need for harmonised annotation principles to improve the comparability, interoperability, and reusability of MWE-annotated syntactic resources.

Two other surveys focus on language resources for MWEs. The first one, by Losnegaard et al. (2016) reports on a survey conducted within the framework of the IC1207 COST Action PARSEME, which aimed to provide a basic overview of the status of various types of language resources across multiple languages. The ultimate goal was to create a portal granting researchers access to previously unavailable or newly developed MWE(-aware) resources. This survey identified approximately 110 resources, including 13 MWE lists, 48 MWE-dedicated or MWE-aware lexicons, such as semantic dictionaries (e.g., wordnets), and 27 corpora and treebanks with annotated MWEs (monolingual, bilingual, and multilingual). The other 19 identified language resources include MWE extraction tools, corpora with annotated MWEs, lists of MWE resources, datasets containing MWEs, annotated texts, pattern dictionaries of verbs, and domain-specific MWE lists. These resources cover 19 languages, including Arabic, Bulgarian, Croatian, Czech, Dutch, English, Estonian, French, German, Hebrew, Italian, Macedonian, Polish, Portuguese, Russian, Serbian, Slovak, Slovene, and Swedish. English was the most represented language with 12 resources, followed by Italian, which appeared in 7 resources. The survey focused on metadata descriptions of the resources highlighting several challenges, including their availability in large language resource repositories and cataloging issues that hinder efficient discovery and reuse by researchers. The discussion and conclusions centered on the standardisation and enhancement of metadata descriptions. However, no definitive set of features for describing MWE lexicons was established. The survey ultimately contributed to the development of a metadata schema for language resources containing MWEs; ultimately, the schema is aimed at facilitating the creation of a meta-resource that could be useful to end users.

The second survey by Markantonatou et al. (2026) builds upon and extends the scope of the previous one, by updating and expanding the inventory of MWE-related resources from 2016 onward, focusing on computational lexicons. This follow-up effort was carried out within the framework of the UniDive COST Action (Savary et al. 2024), specifically as part of Working Group 2, which aims to articulate the lexicon-corpus interface (Barbu Mititelu et al. 2024). The new survey identified 66 resources using strict inclusion criteria with special emphasis on the languages featured in the resources and their level of representation. Thus, the survey is following the UniDive COST Action idea to highlight the extent to which MWE lexicons are developed for less-resourced languages. The survey focuses on key features of computational lexicons that require deeper exploration, aligning with the broader vision of establishing a connected ecosystem

of lexicons and corpora where MWEs are systematically encoded and annotated, ultimately enabling more robust and comprehensive NLP applications. A central goal of the survey was to offer a detailed account of the current state of computational MWE lexicons, while also considering recent developments in the context of LLMs.

From a different perspective, Hubers et al. (2020) report on interdisciplinary work carried out within the framework of the ISLA project, which focused exclusively on idiomatic expressions as a subset of formulaic language. The project aimed to investigate the acquisition and processing of MWEs from the viewpoints of psycholinguistics and second-language acquisition. To support this multi-faceted research, the project developed a set of tools and language resources considered essential. These included (i) a comprehensive database of 374 Dutch idiomatic expressions, annotated with information such as frequency of exposure, familiarity of meaning, usage, transparency, and imageability, along with cross-linguistic data indicating whether similar expressions exist in German and Standard Arabic; (ii) a test instrument designed to evaluate idiom knowledge and to gather data on idiom properties among both native speakers and L2 learners; and (iii) a Computer-Assisted Language Learning (CALL) system, featuring immediate corrective feedback to help second-language learners practice Dutch idioms in a structured environment.

The results showed that both native speakers and second-language learners were capable of reliably assessing idiom properties. However, performance varied depending on factors such as language distance and second-language proficiency.

As future work, the researchers indicated exploring improved methods for identifying MWEs that go beyond simple frequency-based approaches, as well as developing standardised procedures for data collection to ensure comparability across studies. A continuously updated and extensible database of formulaic language resources was also referred as a core outcome of this effort.

3.2 NLP applications and MWEs

Considering a direct relation between MWEs and NLP applications, Constant et al. (2017) provide an overview of how MWEs are handled in these, namely, in parsing and machine translation. Rather than treating MWEs as a purely linguistic phenomenon, their study focuses on MWE processing, defining it as a task with two main sub-tasks: discovery and identification. The former is concerned with finding new MWEs (types) in text corpora and storing them for future use in a repository such as a lexicon. In contrast, MWE identification is defined as

the process of automatically annotating MWEs in running text and classifying them into predefined categories. The study explores the interactions between MWE processing and parsing and machine translation applications, emphasizing how key MWE properties, such as discontiguity, non-compositionality, ambiguity, and variability, present both challenges and opportunities for MWE-aware systems. The language resources considered in this survey include (i) lexical resources, i.e., databases containing MWEs, MWE lexicons, and general-purpose dictionaries containing MWEs; (ii) corpora, i.e., corpora annotated with MWE tags and parallel corpora with sentence-level alignments; and (iii) treebanks, that is, special corpora with syntactic relations annotation where MWEs are explicitly marked. The analysis and discussion of these language resources focus primarily on their relevance for parsing and machine translation.

The authors conclude that, while MWE processing presents challenges, it also offers opportunities. Discontiguity hinders discovery and identification but supports parsing by linking nonadjacent words. The orchestration of tasks is critical for MWE processing, as identification benefits from prior syntactic analysis, while pre-processing is more effective for fixed MWEs like named entities. Additionally, MWE categorisation influences processing strategies, with verbal MWEs requiring different approaches than compounds.

Despite these insights, the authors identify several open issues and research questions. These concern evaluation inconsistencies, which, again, hinder discovery, identification, and re-use, particularly in machine translation. Improved metrics and annotation standards for MWEs were considered needed. Also, comparative studies across tasks, processing strategies, and MWE categories are lacking, limiting insights into optimal processing approaches. Finally, coverage limitations persist, as hand-built MWE dictionaries remain a primary resource but suffer from limited domain and language coverage. Integrating automatic discovery methods into downstream applications (e.g., identification, parsing, and machine translation) could enhance coverage and improve extrinsic evaluation methods.

Ramisch et al. (2023) conduct a review of experiments for MWE identification using data from the SEMEVAL DiMSUM (Schneider et al. 2016) and the three editions of the PARSEME Shared Tasks (Savary et al. 2017, Ramisch et al. 2018, 2020). The study analyzes 40 papers on MWE identification, focusing on often-overlooked experimental factors, such as corpus selection, pre- and post-processing, MWE encoding, evaluation metrics, statistical significance, and error analyses. These aspects significantly influence experimental results and conclusions, underscoring the need for systematic descriptions of experimental conditions to mitigate the risk of misleading findings due to poorly designed se-

tups. The survey aims to base its discussion on quantitative data extracted from the analysed papers, including details on languages covered in the corpora; data partitioning for training, development, and testing; MWE categories targeted by models; corpus pre- and post-processing strategies; MWE encoding and decoding methods, particularly for classification and tagging models; evaluation metrics used; statistical significance of model comparisons; and error analyses conducted. A key focus of the survey is on how unseen MWEs are handled by identification systems. Findings indicate that corpus characteristics, including the number of languages, application domain, and data partitioning strategies, impact experimental outcomes. The authors emphasise that standardised reporting on corpus selection, pre-processing, evaluation metrics, and statistical significance testing is crucial for ensuring the reliability and reproducibility of MWE identification research. Relevant conclusions on open issues include studying automatic evaluation techniques, while regarding the role of manual evaluation, particularly in downstream tasks where MWE quality may influence final outcomes, and new evaluation protocols to address questions such as whether certain MWE categories are more significant than others. The survey ultimately advocates for the gradual adoption of methodological standards and best practices in both shared tasks and independent research, promoting more transparent, replicable, and robust experimental frameworks for MWE identification.

More recently, Mahajan et al. (2024) present a survey on the approaches, techniques, and features required for MWE identification, presenting language-specific descriptions. The study includes a detailed list of tools used for the extraction and identification of MWEs, such as parsers, semantic analyzers, chunkers, part-of-speech (PoS) taggers, word aligners, and ontology editors. It offers a literature survey on methods and techniques employed for MWE processing in non-Indian languages, covering seven European languages, three Asian languages, and three African and Middle Eastern languages. The survey also includes a dedicated section that reviews the methods and techniques for MWE identification for seven Indian languages. The work draws attention to recent progress in applying machine learning techniques to Indian languages while also identifying major challenges, including the lack of annotated corpora, the diversity of writing styles, and the complexity of morphological structures. By examining various MWE identification approaches – linguistic, statistical, machine learning, and hybrid – particularly in the context of Indian languages, it underscores the importance of combining methodologies to improve accuracy. Moreover, the study highlights a significant gap in research on MWEs in some Indian languages, such as Marathi, emphasizing the need for further investigation and for the development of specialized approaches tailored to these languages.

Unlike the surveys discussed in this section, the present chapter focuses on MWE lexical resources, integrating both perspectives: the description and annotation of MWEs, and the use and contributions of these resources to NLP tasks and applications. Concentrating on recent developments, we aim at identifying MWE lexical resources through a comprehensive literature survey, further complemented by a community-wide questionnaire. The resources examined include lexicons specifically dedicated to MWEs, as well as more general datasets that incorporate MWE information.

We provide a detailed analysis of these lexicons, considering multiple dimensions such as lexicon type, categories of expressions included, language coverage, and the linguistic information encoded. These aspects are also examined in relation to their use and impact on NLP tasks and applications, including their relevance to recent advances in LLMs and emerging directions in this area of research and development.

By offering an in-depth analysis of related work, along with a systematic literature review and a dedicated community questionnaire, this study goes beyond the description of MWE representation in language resources and their handling in NLP applications. It aims to assess the broader impact of MWE lexicons on NLP tools and applications, while identifying critical gaps and emerging trends in the field.

4 Methodology

This section provides a detailed overview of the methodology used in our study, describing the processes for selecting the relevant papers and for developing a questionnaire on the use of MWE lexical resources by the community (including both academic and industrial environments).

4.1 Paper selection process

The paper selection process followed a structured approach. It began with an initial search and filtering of papers based on relevance and publication date. This was followed by peer review by team members. Finally, a list of selected articles was compiled, forming the basis for the analysis in this survey.

To curate a selection of relevant papers for our survey, we focused on three prominent sources: the ACL Anthology,[1] EUROPHRAS,[2] and arXiv.[3] To identify

[1]https://aclanthology.org/, last accessed in April 2025
[2]https://europhras.org/conferences/, last accessed in April 2025
[3]https://arxiv.org/, last accessed in April 2025

relevant papers from the ACL Anthology, we used an automated search method using carefully selected keywords related to MWEs, including *idioms*, *collocations*, and *phrasal verbs* as well as their application in NLP tasks such as *identification*, *parsing*, *machine translation*, etc. This automated search method allowed us to efficiently scan a large number of publications and identify those that were most relevant to our survey. In contrast, the EUROPHRAS and arXiv selection process involved the manual extraction of papers. In addition, after the initial search, reference lists of recent papers were reviewed to identify new and relevant papers for inclusion. To ensure the inclusion of recent and relevant research, papers published within the last six years (January 2019 to March 2024) were considered. A total of 68 papers were retrieved.

After the initial search and filtering, each extracted article was reviewed by two team members to assess its relevance to the survey. In cases where the two reviewers disagreed on the relevance or inclusion of a paper, a third member of the team was brought in to read the paper and make the final decision. All results of this review process were then discussed in regular meetings, allowing consensus to be achieved on the final selection of articles. This approach reduced the risk of omitting important papers or including less relevant ones.

The final list of 37 papers was compiled and served as the basis for the analysis presented in the survey (Section 5).

4.2 Devising the questionnaire for the community

To gather detailed information about current practices and the use of MWE lexicons in various NLP tasks, in addition to the literature review mentioned above, a questionnaire was created for the user community. The primary objectives of the questionnaire are as follows:

- to understand the professional demographics and activities of individuals working with MWE lexicons;
- to collect detailed information about the use, characteristics, and effectiveness of MWE lexicons in NLP tasks;
- to identify specific challenges and requirements associated with the use of MWE lexicons;
- to gather information about the supplementary resources and integration guidelines for MWE lexicons (if any) in NLP systems.

The questionnaire design was grounded on several interdisciplinary approaches within the fields of computational linguistics, NLP, and lexicography. Generally, it is focused on the identification of the specialized resources used to improve syntactic parsing, semantic analysis, and machine translation; the techniques for automatic extraction and application of MWEs; and the lexicons' characteristics such as size, format, update frequency, etc. used for the above-mentioned tasks.

The questionnaire was structured to include the following sections:

- **Participant consent and understanding**. This section was developed with the purpose of ensuring the informed and voluntary consent of the participants. The questions confirm that the participants understand the purpose of the survey and agree to the terms, ensuring ethical compliance and data protection.

- **Professional background**. On the one hand, questions on the area of professional activities, roles or functions, and community representation help identify the demographic and professional distribution, which is important for data segmentation and targeted analysis. On the other hand, these aid in understanding how the professional environment might influence the participants' experiences and reveal challenges faced during the use of the lexicons.

- **Experience with MWE lexicons**. This section gathers information on participants' experience with MWE lexicons over the last five years, focusing on practical applications and specific NLP tasks. Understanding the diversity and depth of experience makes it possible to assess the impact of MWE lexicons on different applications.

- **MWE lexicons characteristics**. This section collects specific details about the lexicons' use, including their title, size, format, type, direction, language coverage, access, and update frequency. These detailed attributes are important for evaluating different MWE lexicons.

- **Perception on MWE lexicons use in NLP**. This section focuses on questions about specific features, complementarity with other resources, challenges, and evaluation methods, as well as questions about guidelines, community-driven approaches to integrating MWE lexicons in NLP systems, and recommendations help identify common problems and best practices.

5 Critical literature review on MWE lexicons and their contribution to NLP tasks and applications

In this section, we present the results obtained from the critical literature review focusing exclusively on lexical resources reported to have been used in an NLP downstream task or application. Based on the papers selected, we want to shed light on the following aspects:

- papers identified and the venues in which they were presented;
- lexicons reported; we elaborate on the features of the lexicons (linguality type, languages covered, types of MWEs treated, and linguistic information that is encoded for MWEs);
- NLP task or application for which the MWE lexicon is used; the focus is on the linguistic information that was used in the task and the overall impact of incorporating/using the MWE resource;
- the interplay between LLMs and MWEs.

The papers' analysis grid is provided as supplementary material accompanying this chapter and is openly available via Zenodo (Giouli et al. 2026c).

5.1 Papers' profile

The paper selection process resulted in a total of 37 papers, originating mainly from conference proceedings (34 conference papers, 1 journal paper, and 2 arXiv preprints).

Most of the publishing venues were related to computation and NLP, although venues focused on language resources were also relevant, as shown in Figure 1.

As expected, many of the analysed papers originated from MWE-dedicated venues (13), while the remainder were fairly evenly distributed across various other events (see Figure 2). Regarding publication dates, most papers were published in 2022, followed by 2024 and 2023, with several editions of the Workshop on MWEs contributing to these numbers, as illustrated in Figure 3.

5.2 Lexicons and their features

We present here several characteristics of the MWE lexicons used in NLP tasks, such as the languages covered and the type of MWEs they describe, as well as the linguistic information encoded for the contained MWEs. The complete list

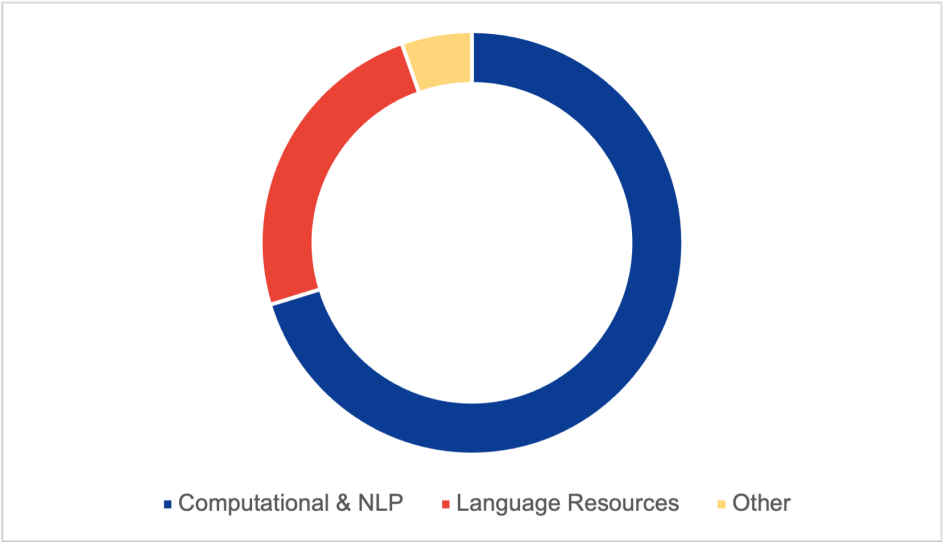

Figure 1: Area of the publishing venues

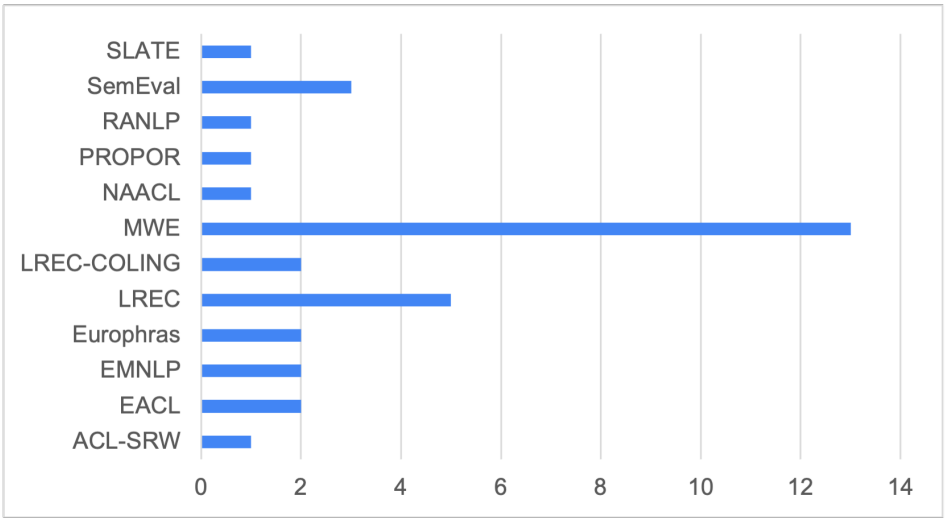

Figure 2: Paper distribution per conference

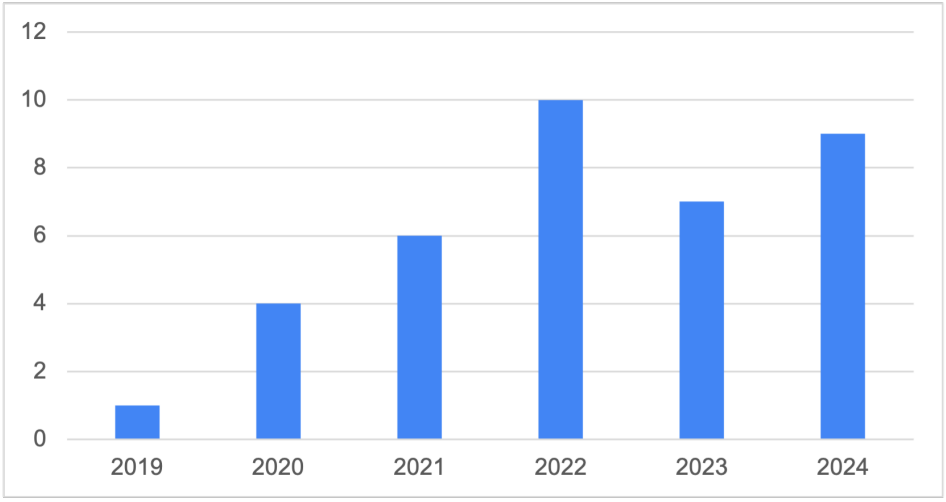

Figure 3: Papers' year of publication

of resources analysed is provided as supplementary material accompanying this chapter and is openly available via Zenodo (Giouli et al. 2026a).

5.2.1 Languages

The analysed lexical resources represent monolingual, bilingual, and multilingual MWE lexicons that cover a variety of languages or language varieties: English, Russian, French, German, Dutch, Italian, Portuguese, Spanish, Galician, Polish, Czech, Belarusian, Bulgarian, Urdu, Chinese, and Arabic.

30 out of the 37 surveyed papers discuss monolingual resources, 3 papers are devoted to experiments on bilingual resources, and 4 articles deal with multilingual data. The distribution of languages per resource is presented in Table 1.

Table 1: Number of languages in the resources

	monolingual	bilingual	multilingual
No.	30	3	4

5.2.2 Types of MWEs included in the lexicons

The lexical resources mentioned range from computational lexicons (11) and wordnets[4] (3) to datasets[5] (29) designed for MWE identification. They include n-grams, terms (2 resources), idioms (sometimes with paraphrases), lists of MWEs and their frequencies in corpora, but also datasets with compositionality ratings for MWEs.

Of the 43 lexical resources surveyed, 25 are dedicated to MWEs only, focusing on the detailed structure and semantics of MWEs, while the remaining 18 are MWE-aware. [6]. In the latter case, MWEs are integrated into a wider range of single- and multiword entries.

The MWEs in the different resources can be classified into several types based on their grammatical and functional characteristics, namely verbal MWEs, nominal MWEs, adjectival MWEs, adverbial MWEs, or functional MWEs. From another perspective, collocations, idioms, phrasal verbs, proverbs, terms, etc. form a typology of MWEs with unique linguistic properties. Figures 4 and 5 show the distribution of different types of MWEs across lexical resources. Figure 4 presents MWEs classified based on grammatical and functional characteristics. The "All" label indicates lexicons that provide broad coverage across all MWE categories. In contrast, Figure 5 illustrates a typology grounded in lexical and idiomatic properties: proverbs, terminological terms, phrasal verbs, collocations, and idioms.

As shown in Figure 4, resources that categorize MWEs based on grammatical and functional characteristics tend to emphasise nominal and verbal types, which are predominant in the majority of the lexicons examined. Adverbial and adjectival MWEs are less represented. The inclusion of an "All" category suggests that certain resources are designed to provide comprehensive coverage of all MWE types. In contrast, Figure 5 shows that MWEs categorised by lexical and idiomatic criteria, such as idioms, collocations, phrasal verbs, terms, and proverbs,

[4]A wordnet is any resource that follows the organisation principles of Princeton WordNet (Miller et al. 1990): a lexical-semantic network whose nodes contain synonym sets which are connected by semantic relations; verbs and nouns are hierarchically organized according to hyponymy/hyperonymy, meronymy/holonymy, troponymy relations, descriptive adjectives are organized in pairs of direct antonyms with adjectives in a semantic similarity relation clustered around each of the members of the respective pair; adverbs and relational adjectives have no hierarchical organisation.

[5]We use the term *dataset* to refer to a mere inventory of linguistic phenomena, here of MWEs.

[6]All lexical resources used in the surveyed papers are identified independently, regardless of (i) being used in more than one paper; (ii) being used in conjunction in only one paper. This means that the list contains 43 lexical resources and not 37 as shown in Giouli et al. (2026a)

are generally less represented. Idioms and collocations are the most common, while proverbs are particularly rare.

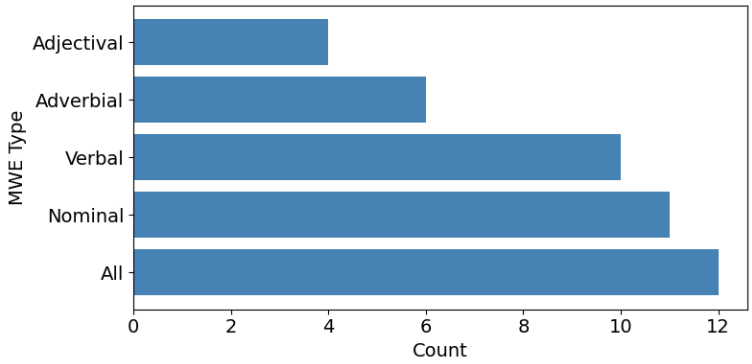

Figure 4: Distribution of MWE types across the lexical resources (N = 43), based on grammatical and functional classifications

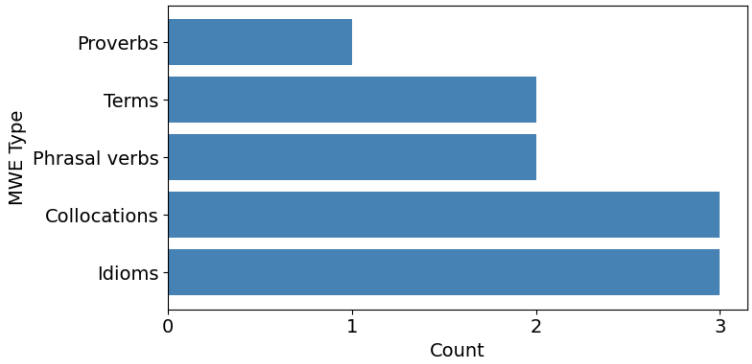

Figure 5: Distribution of MWE types across the computational lexicons surveyed (N = 11) based on lexical and idiomatic classifications

5.2.3 Types of linguistic information encoded about MWEs

The lexicons we identified in this survey typically encode MWE-related information beyond the one used in the task reported. The nature and extent of the linguistic information encoded depend on the type of the lexical resource (whether MWE-aware or MWE-dedicated), its intended purpose, and the overarching framework (linguistic, task- and application-oriented, etc.). This information is subsequently utilised in NLP tasks based on the specific requirements

of the task at hand. The diversity of approaches to MWE encoding in lexicons should be noted, as it poses challenges for any systematic analysis.

In MWE-aware lexicons, such as wordnets, information about the lemma and part-of-speech is typically encoded. In such resources, MWEs are treated following the same principles as single-word entries. On the contrary, MWE-dedicated lexical resources primarily encode part-of-speech information and the type of MWE, such as idioms, light verb constructions, verbal collocations, and compound nouns (see Section 5.2.2). This classification supports a more nuanced understanding of MWEs and aids in their proper treatment.

Various MWE-dedicated lexicons encode syntactic information, often encompassing internal structure (Baptista 2012), argument structure, and syntactic dependencies between their constituents (Espinosa Anke et al. 2019, Müller et al. 2022), or even distributional constraints on argument slots and transformational properties (Baptista 2012). Such information is aimed at modeling the behavior of MWEs in context in a straightforward and explicit manner. In contrast, other resources provide sample sentences that depict the distribution of MWEs in an implicit way (Cook et al. 2008, Haagsma et al. 2018, Tayyar Madabushi et al. 2021, Zhou et al. 2021).

As mentioned above, three wordnets were identified in the study, targeting English (Fellbaum 2010), Polish and English (Rudnicka et al. 2015), and German (Henrich & Hinrichs 2011). These resources typically encode semantic information by structuring both single- and multiword lexemes into synonym sets (synsets), which are accompanied with glosses (definitions). In addition, they capture lexical-semantic relations and establish interlingual correspondences via the Inter-Lingual Index (ILI). Besides glosses, paraphrases are often used to represent the semantics of MWE entries in lexicons or lexical datasets (Tayyar Madabushi et al. 2021, Liu & Hwa 2016, Pershina et al. 2015). More elaborate MWE-dedicated datasets encode sense information for MWEs using fine- or coarse-grained sets of meanings along with usage information (Tayyar Madabushi et al. 2021). Moreover, lexical functions, as introduced in the Meaning-Text Theory (Mel'čuk 2007), are also used as a device to capture the semantics of collocations in Espinosa Anke et al. (2019).

In semantics-oriented tasks, such as sentiment analysis, the identification and interpretation of MWEs play a crucial role. In the survey, specialised MWE-aware lexicons tailored to meet the needs of such applications were identified; in these lexicons, both single- and multiword entries are labeled for semantic orientation and intensity (Moreno-Ortiz et al. 2020). Reflecting the specific demands of sentiment analysis, some of the MWE-dedicated lexicons in our survey are designed to encode discrete emotion categories (Prochnow et al. 2024), while others

focus on capturing prior and contextual sentiment polarity: positive, negative, or neutral sentiment (Jochim et al. n.d., Moreno-Ortiz et al. 2020).

Certain datasets identified in the survey specifically target the so-called possibly idiomatic expressions (PIEs), that is, expressions that can be interpreted either literally or idiomatically depending on context (Haagsma et al. 2018, Zhou et al. 2021, Cook et al. 2008). These datasets provide contextualized examples accompanied by fine-grained annotations, including binary or categorical labels indicating whether each instance is used literally or idiomatically. Similarly, additional semantic aspects are captured through ratings for idiomaticity, semantic opacity, and familiarity degree (Hubers et al. 2020).

The extent to which a lexicon incorporates linguistic information typically reflects the specific requirements of the target task. In addition to core lexical data, some resources encode supplementary – or domain-specific – linguistic information, such as language proficiency levels (Capel 2015, de Jong et al. 2016). Others provide links to external resources (Collard et al. 2023), or offer only minimal lexical detail (Sarwar 2022), indicating a more diverse or limited scope.

In summary, the lexicons used in NLP tasks can vary significantly in terms of the type of linguistic information they encode. While some lexicons contain only linguistic features relevant to a specific task (e.g., PoS, syntactic, and semantic information), others incorporate different kinds of information that support a broader range of applications (see Table 2).

Table 2: Typical layers of linguistic information encoded in MWE lexicons.

Information type	Fields / tags	Found in
Form	lemma, PoS, inflection type	MWE-aware and MWE-dedicated lexicons
MWE type	idiom, LVC, collocation, compound	MWE-dedicated lexicons
Syntax	internal modifier, word-order permutation, verb alternations	MWE-dedicated lexicons
Semantics	gloss, paraphrase, sense	wordnets and task-oriented lexicons
Usage	frequency, corpus examples, language-proficiency level	task-oriented lexicons

Notably, MWE-specific challenges, particularly those concerning the assignment of lemmas and canonical forms, word order, modification, and non-contiguous elements, are largely overlooked in the surveyed lexicons.

5.3 MWE lexicons in NLP tasks and linguistic information used

We provide an overview of the NLP tasks for which the use of a MWE lexicon has been reported, highlighting the type(s) of information provided by such resources that proved relevant or useful for these tasks.

5.3.1 NLP tasks for which a MWE lexicon is used

MWE lexicons are used for NLP tasks that vary from MWE detection and extraction, bilingual term extraction, lexicon development by extraction, lexicon induction, sentiment labeling, machine translation, neural machine translation to modeling, recognizing, and categorizing MWEs, testing idiomaticity and improving language models to capture idiomaticity, natural language parsing, language generation, and evaluation of LLM performance on MWEs. The tasks also include gender identification, idiom and MWE teaching and learning systems, idiomatic compound detection and semantic relations for terminology, including synonymy, antonymy, and hypernymy.

Generally, a notable subset of papers focuses specifically on idiomatic expressions, indicating a significant interest in understanding their complexities. For instance, Lee & Uvaliyev (2023) examine the selection of MWEs for graded vocabulary lists, emphasizing the role of phrasal verbs and idioms in language learning, while other articles apply a more general approach to idiomatic expressions. At the same time, some other papers have a general focus on MWEs, analysing their contextual information, structure, and semantic and syntactic idiosyncrasies.

The NLP task for which MWE lexicons are most frequently used is MWE identification, with 21 out of 37 papers reporting this use (i.e., 56%). Given that (multiword) terms and named entities can also be treated as MWEs, this number could actually be increased by two more papers (thus accounting for 62% of all the literature reviewed). Idiomaticity is directly addressed in two papers, but if we include literature dealing broadly with the identification of concepts for which a single word does not exist in the target language – and which are instead expressed through multiword combinations – then we arrive at a total of 25 papers (corresponding to 67%) that relate to idiomaticity in a broad sense.

Another three papers (8%) report the use of MWE lexicons for sentiment labeling. Given that MWEs are, more often than not, associated with a connotation, be it positive or negative, this explains the interest in MWEs for this downstream task.

Other tasks, each illustrated by one paper in our survey set, are: sentence simplification, sentence generation, author gender identification, understanding

how LLMs work and probing them, lexical complexity, validation of lexicon or grammar, and machine translation.

5.3.2 Linguistic information for MWE processing in NLP tasks

The type of linguistic information revealed by the studies analysed includes several crucial elements that contribute to the understanding and processing of MWEs, and includes a diverse range of MWEs such as idioms, phrasal verbs (Lee & Uvaliyev 2023, Zhou et al. 2021 and others), and nominal phrases (Maziarz et al. 2022, Premasiri et al. 2022 and others). Key components encompass annotated dependency trees, including the representation of syntactic structures, where words are connected to their syntactic heads, showing how each word in a sentence relates to others, part of speech tags used to assign grammatical categories to words, identifying the roles of different words within an expression, and facilitating the analysis of their roles in MWEs (Collard et al. 2023). Surveys on the semantic relations between constituents explore how these expressions interact within contexts (Fisas et al. 2020, Maziarz et al. 2022 and others). Additionally, lemmatisation, which is the process of reducing words to their base forms, allows for a more consistent analysis across different word forms.

The linguistic information used in these studies describes various aspects that are essential for analysing MWEs. This includes a variety of lexical features, such as morphological and syntactic characteristics that are important for understanding how MWEs function within the language. The resources often capture information regarding the meaning and interpretation of the expressions, emphasizing their non-compositionality and context-dependency. Furthermore, the frequency of use (e.g. Hubers et al. 2020) plays a critical role, as it affects how MWEs are processed in different linguistic contexts.

A key area of focus is the semantic and syntactic classification of MWEs. For example, Zhou et al. (2021) focus on the semantic classification of idiomatic expressions, specifically examining their compositionality. This classification is crucial for tasks such as paraphrasing and idiomatic sentence generation, as it helps determine how these expressions can be effectively utilized and recognised in various contexts. In another study, Espinosa-Anke et al. (2021) address the classification of collocations, where MWEs are classified based on their compositional or semi-compositional nature. This classification improves the understanding of MWEs in natural language processing tasks such as lexical resource development and language modeling.

When addressing sentiment labeling, a mere list of MWEs will prove insufficient; thus, resources containing emotion annotation, polarity, and intensity of MWEs are rather useful.

5.3.3 Methodologies employed

The experiments used in the papers analysed include different tasks and methods, some of which were not directly dependent on the lexicon development, but provided their integration into various NLP tasks. One of the main goals of these tasks is the identification and extraction of MWE, such as idiomatic expressions, collocations, and phrasal verbs from text (Tanner & Hoffman 2023), and for this purpose, the lexicons serve as structured resources for recognizing MWEs in corpora. As a result, various techniques are applied, including rule-based systems, which rely on predefined linguistic rules to identify MWEs based on patterns in the text (Maziarz et al. 2022), machine learning methods, where models are trained on annotated data to learn how to identify MWEs (Kanclerz & Piasecki 2022), and statistical processing used to find or identify MWEs in large datasets. For example, some techniques focused on identifying MWEs through predefined patterns, which aid in the recognition of nonliteral expressions, and for this purpose, the PARSEME corpora were applied (Müller et al. 2022).

Furthermore, researchers generally address three MWE-related tasks, especially, evaluating and improving automatically generated MWE resources (Lee & Uvaliyev 2023), curating lexicons for Machine Translation (MT) systems to cover idioms and other MWEs (Zhou et al. 2021) and detecting idiomaticity in context, often in zero- or one-shot settings (Jakhotiya et al. n.d.). Studies use a range of techniques and methods like transformer fine-tuning, where models such as eBERT-multilingual-uncased, DistilBERT-multilingual-uncased, XLM-RoBERTa-large, and XLM-Net are trained to classify tokens as idiomatic (Jakhotiya et al. n.d.); AMR-driven chain-of-thought method that prompts Abstract Meaning Representations (AMRs) to guide LLMs in detection of idioms (Jin et al. 2024); gathering synsets to improve MWE recognition accuracy (Maziarz et al. 2022). In any case, the resulting systems are evaluated by fine-tuning and testing on zero- and one-shot settings to evaluate how well transformer models can detect idiomatic expressions or classify MWEs (Jakhotiya et al. n.d.). Such evaluations help refine the models and ensure that the MWE lexicons are contributing meaningfully to the task.

In addition, the lexicons that include sentiment-related annotations, such as polarity and intensity, are used to improve the accuracy of sentiment analysis models. As a result, two main tasks are revealed, especially those focused on sentiment evaluation (Hwang & Hidey 2019), in which MWE-aware resources help models to label expressions and, those dedicated to the acquisition and learning of idioms (Hubers et al. 2020), where a combination of linguistic, educational, and other perspectives guide models how idiomatic meaning is created.

In general, the methodologies involved demonstrate the diverse strategies for MWE detection, classification, and integration into NLP applications. And different experiments demonstrate that MWE lexicons play an important role in various NLP tasks, both as basic resources for rule-based and statistical approaches and as components for training and evaluating machine learning and transformer-based models.

5.4 MWEs and LLMs

In general, the integration of MWEs into the pre-training or fine-tuning stages of LLMs still remains limited. Despite the recognized importance of MWEs for various NLP tasks, there are relatively few efforts that systematically incorporate MWE lexicons into LLM pipelines. This suggests a gap in current practices, where the potential of MWE-aware or MWE-dedicated lexical resources remains largely unexplored. In the next sections, we focus on the impact of MWE lexicons on the pre-training and fine-tuning of LLMs, and on the mutual impact between MWE datasets and LLMs.

5.4.1 Computing multiwords for LLMs

Byte-Pair Encoding (BPE) tokenisation, as introduced by Sennrich et al. (2016), is an adaptation of the original BPE data compression algorithm (Gage 1994) for subword segmentation in NLP. This method has since been developed into BPE variants tailored for MWEs, such as those incorporating pointwise mutual information to guide merge operations (Kumar & Thawani 2022). Another approach, which represents statistically determined MWEs as individual tokens, is used to compress the text input of LLMs (Gee et al. 2023). The Principle of Least Effort, where the tokeniser independently learns an integrated vocabulary of subwords, words and MWEs, has also been introduced (Yang 2024).

Kumar & Thawani (2022) investigate the incorporation of MWE-based tokenisation into transformer-based neural machine translation systems. At the same time, the architectures and techniques they examine can be directly applied to LLMs, which typically employ BPE or similar subword tokenisers. In contrast, Yang (2024) explicitly targets tokeniser design for LLMs by introducing the Less-is-Better (LiB) model, a cognitively inspired tokeniser that learns a joint vocabulary of subwords, words and MWEs. The LiB model aims to minimise both the number of tokens and vocabulary size.

MWE-based tokenisation is not widely used in LLMs, mainly due to the challenges of integrating fixed multiword units into dynamic subword-based models

and the lack of large-scale, high-quality MWE resources suitable for training. It has only been experimentally investigated to a limited extent (mainly for machine translation). Such approaches recognise the value of MWEs as language units, although their main goal is to reduce computational costs.

Alternative recent approaches build on the idea for multitokens being directly included in the algorithms for pre-training the LLMs. In Multitoken Prediction (MTP), LLMs are equipped with some additional parameters to predict multiple tokens simultaneously and in parallel using independent output heads (Gloeckle et al. 2024). DeepSeek V3 (DeepSeek-AI et al. 2025) has modified the MTP so that additional tokens are predicted sequentially, and the full causal chain is preserved at each prediction depth. However, MTP still remains experimental and has not yet become widely adopted in mainstream LLM architectures.

5.4.2 Fine-tuning of LLMs for the recognition of MWEs

Several experiments have investigated how LLMs can be fine-tuned to improve their performance in MWE classification. Fine-tuning of pre-trained models requires annotated data suitable for a classification task. In the broadest sense, datasets containing annotated MWEs can be exploited as lexicons, as the annotated multiword units can be easily extracted from the datasets and classified based on the annotated attributes and their values.

The following language models were mainly used for the fine-tuning experiments: BERT (base and large, uncased) (Devlin et al. 2019), RoBERTa (base and large) (Liu et al. 2019), DistilBERT (Sanh et al. 2019), ALBERT (Lan et al. 2020), XLNet (base and large) (Yang et al. 2019), and XLM-RoBERTa (Conneau et al. 2020).

For example, BERT, DistilBERT, XLM-RoBERTa, and XL-NET are the models used in an experiment to recognise idiomatic expressions using binary classification in two different settings: zero-shot and one-shot (Jakhotiya et al. n.d.). The training and development dataset (Tayyar Madabushi et al. 2021) consists of samples containing one target sentence, one MWE (with or without idiomatic meaning), and two context sentences. The results show that the one-shot setting performs better than the zero-shot setting.

In many cases, fine-tuning of LLMs is conducted for a single language. For example, using the Romanian annotation in the multilingual PARSEME corpus for several types of VMWEs (Ramisch et al. 2020), significant improvements were achieved in the identification of unseen Romanian MWEs (Avram et al. 2023) by using a lateral inhibition layer and adversarial training for multilingual LLMs such as XLM-RoBERTa.

The impact of Pattern Exploit Training (Schick & Schütze 2021), a few-shot method of classification, and BERTRAM (Schick & Schütze 2020), a method of creating contextual embeddings, is evaluated on the task of idiomaticity detection. In evaluating the models, SemEval 2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al. 2022) was used. The dataset consists of three languages: English, Portuguese, and Galician. The results suggest that few-shot learning methods are effective on the task of idiomaticity detection despite the lower accuracy on Portuguese and Galician. Similarly, the results support the conclusion that improved MWE representations does have an impact on improved detection.

Three experiments are carried out for French (Bui & Savary 2024): (i) fine-tuning the models on Sequoia (Candito et al. 2020), a French corpus annotated for all MWE types; (ii) treating verbal (PARSEME dataset for French) and non-verbal MWEs as separate tasks (union approach), training the models on the respective datasets and then merging the predictions; (iii) a multi-task approach in which the models are trained jointly on both types (verbal and non-verbal). The results show that the combination of independent models (union approach) performs better than single-task and multi-task models, especially when using monolingual transformer models: CamemBERT (Martin et al. 2020) and FlauBERT (Le et al. 2020). The conclusion was that the union approach effectively utilises the complementary strengths of the models trained on different datasets and provides a robust solution for MWE identification.

To summarise, various fine-tuning methods are used in BERT-based transformers to identify MWEs. Fine-tuning requires annotated MWE datasets instead of lexicons. However, as discussed above, these annotated datasets can be considered some sort of lexicons because they encode structured linguistic information relevant to MWE processing.

5.4.3 Evaluation of LLMs for the recognition of MWEs

Studies on encoder-only models have shown that pre-trained language models do not adequately capture and represent the meanings of idiomatic expressions (Garcia et al. 2021). A set of probing measures and a dataset for English and Portuguese are proposed to investigate how accurately idiomaticity in MWEs, especially in noun compounds, is captured in vector space models, focusing on some widely used representations such as BERTRAM, BERT, SBERT, DistilB, ELMo (Peters et al. 2018), and GloVe (Pennington et al. 2014). The probing tasks have shown that contextualised models may not recognise idiomatic noun compounds

that have a lower degree of substitutability of the individual components compared to more compositional noun compounds.

It was shown that all transformer models outperform the Long Short-Term Memory (LSTM)-based method in recognising MWEs, and it was concluded that transformer models can overcome the challenges posed by MWEs in local domains such as plant names and flower names better than previous neural methods such as LSTM (Premasiri et al. 2023).

Furthermore, generative models such as GPT-2 perform poorly when processing figurative language without the use of mitigation strategies (Jhamtani et al. 2021). A survey of recent work on MWE semantics has shown that, while transformer-based language models with an encoder-only architecture (BERT, mBERT, XLM-RoBERTa, RoBERTa, ALBERT, DistilBERT, SBERT, DeBERTa) and GPT capture some aspects of MWE meaning, this ability is neither comprehensive nor consistent (Miletić & Schulte im Walde 2024). The conclusion was that improvement with optimisation strategies such as fine-tuning and knowledge expansion would yield very different gains. This can be explained by the fact that MWE representations rely on memorized information rather than sophisticated meaning processing, which is reflected in the suboptimal generalisation ability. Previous research has shown that encoder-only pretrained language models can effectively represent MWEs when fine-tuned with appropriate data (Tayyar Madabushi et al. 2022).

Experiments were conducted on an English language dataset to evaluate the ability of LLMs such as the 7-billion-parameter (7B) versions of Llama 2, Vicuna, and Mistral to detect idiomatic expressions in a zero-shot setting (De Luca Fornaciari et al. 2024). The results of the study show that when it comes to recognising the meaning of an ambiguous sentence, LLMs struggle to distinguish between literal and idiomatic language use. Further experiments on the recognition of MWEs were carried out in English, Portuguese, and Galician with different generative models from the GPT family, the Gemini family, the Llama 2 family, the Flan-T5 family, Phi-2, and Mistral (Phelps et al. 2024). The results showed that generative language models deliver competitive results, but do not reach the performance of fine-tuned, pre-trained language models.

Further experiments were conducted with open-source and proprietary LLMs (from the GPT family and the Lamma 3 family) to investigate how these models handle MWEs in different languages and contexts, using a dataset with examples in three language pairs – Galician-Spanish, English-Spanish, and Portuguese-Spanish (De Leon et al. 2025). The ability of LLMs to recognise whether MWEs are used as idioms or literally, the extent to which LLMs effectively capture the non-compositional meanings of MWEs, the ability of LLMs to acquire

information from the prompt, and the relationship between LLM size and performance on texts containing MWEs with non-compositional meanings were investigated. The experiments show that LLMs, despite their abilities in many other tasks, continue to have difficulties processing MWEs. The conclusion was that this gap highlights the limitations of these models compared to fine-tuned language models such as XLM-RoBERTa, which are much smaller models and contain far fewer parameters. While some results show that the processing of MWEs with non-compositional meanings in English is comparable to fine-tuned pre-trained language models, the same cannot be said for other languages.

5.5 Impact of using MWE lexicons

The current survey reveals that many NLP tasks are idiom-related. MWE lexicons are used to extract idioms (Maziarz et al. 2022) from corpora, to detect (the degree of) idiomaticity (Jakhotiya et al. n.d.), and enhance idiom learning approaches (Krotova et al. 2020). Lu (2022) uses a lexical resource to improve the handling of MWEs in ML, specifically for classifying idiomatic expressions. In the work of Liu & Lareau (2024), the model correctly guesses the masked token around 60% of the time for tokens within idioms, compared to only 25% for simple lexemes. There is no significant difference between the three types of idioms: 62% for weak idioms, 58% for semi-idioms and 62% for strong idioms. This reveals again the model's higher capacity in predicting tokens within idioms than simple lexemes.

The impact of incorporating MWE resources into linguistic tasks is significant, particularly in improving the performance of applications such as translation, vocabulary extraction, and language generation. For example, Lee & Uvaliyev (2023) report that corpus-based evaluations of automatically generated vocabulary lists benefited from using MWEs. Especially, the use of MWEs allowed for an evaluation of semantic compositionality, which led to more accurate vocabulary extraction. Kumar & Thawani (2022) discuss the impact of MWEs on neural machine translation, particularly between English and Hindi. Their findings indicate that applying naive extensions of BPE to MWEs resulted in suboptimal performance, as the method often split MWEs incorrectly, distorting their meanings. The research highlights the need for more sophisticated methods that account for the specific properties of MWEs in translation tasks. These and other findings demonstrate that the proper integration of MWE resources can greatly enhance both the accuracy and contextual relevance of linguistic tasks.

The study by Haviv et al. (n.d.) shows that idioms help us understand how transformer models make accurate predictions. The authors state that idioms are useful for studying memorisation. MWE lexicons are also used for idiomatic

sentence generation and paraphrasing (Zhou et al. 2021). There are also experiments on sentence similarity identification carried out by Knietaite et al. (2024) that explore the impact of using the local context and glosses for MWE component words, revealing that it is beneficial for the sentence similarity task. The authors observe an increase in performance as the dataset size increases and the impact of using glosses. However, the authors identify that it does not improve the models' representation of idioms.

There is an impactful part of research effort focusing on MWE lexicons and datasets related to general MWE identification and extraction by working on training and testing the datasets (Premasiri et al. 2022, Maziarz et al. 2022). Mohamed et al. (2024) advocate that lexicons even with minimal morphosyntactic information, i.e., PoS information and verbal MWE types, can amplify current MWE-annotated datasets and refine identification strategies. Müller et al. (2024) demonstrate that lexicons containing broader lexical coverage, i.e., more detailed syntactic description, particularly about distinguishing between sentence-internal and sentence-external modifiers, facilitate accurate parsing of MWEs. Such resources can also amplify current MWE-annotated datasets, as well as improve identification strategies.

NLP tasks also involve lexicons for enhancing sentiment analysis. For example, Negi et al. (n.d.) show that using MWE resources enhances the model's ability to accurately extract and classify aspect terms, provide more context-aware sentiment analysis, and improve performance in domain-specific scenarios where MWEs play a significant role. The integration of MWEs into Aspect-Based Sentiment Analysis pipelines ensures a more sophisticated analysis. The use of lexicons also improves alignment and translation quality in NLP applications (Adjali et al. n.d.).

The use of a MWE lexicon enhances automatic author gender identification as noted by Sarwar (2022). The author notes that in experiments using only single words as features, only MWEs as features, and MWEs plus single words as features, the last option seems to perform better for most classifiers (i.e., Support Verctor Machine (SVM) performs best) and pre-trained models (Multilingual BERT, DistilBERT, XML-RoBERTa, Multilingual DeBERTa). However, the experiments in Structural Topic Modelling (STM) reveal a twofold insight (Kokkinakis et al. 2023). The results suggest that the STM models perform only slightly better without MWEs than with MWEs. One reason could be that the vocabulary is larger without MWEs included, so the number of topics captures more words. Still, there is some indication that MWE identification leads to better interpretability of the STM, as calculated by the semantic coherence, which was

higher for the dataset with MWEs. Therefore, it can be concluded that the identification of MWEs can slightly enhance the explainability of the generated topics and themes, which could lead to a more appropriate labeling of the topic itself during the qualitative interpretation of the generated topics, i.e., incorporating MWEs into the models creates slightly more informative resulting topics.

6 Users' survey

This section focuses on the user survey, providing information on the questionnaire's dissemination, respondent profiles, data and results obtained, and an analysis of the comments and suggestions collected. The goal is to understand how MWE lexicons are used in research and language technologies, and how this knowledge can inform and foster synergies between lexicography and NLP, particularly in relation to MWE identification and discovery.

6.1 Dissemination and respondents' profile

The survey was implemented using Qualtrics[7] and distributed via email through an open-access link. The final questionnaire was shared across several mailing lists and scientific communities, including: the Distant Reading COST Action,[8] the European Lexicographic Infrastructure (ELEXIS)[9] community, the European Network of e-Lexicography (ENeL),[10] the European Association of Lexicography (EURALEX),[11] Keystone – Semantic Keyword-Based Search on Structured Data Sources Cost Action,[12] Parsing and Multiword Expressions (PARSEME) scientific network,[13] Processing of Portuguese (PROPOR) community, and Universality and Diversity in Language Technology (UniDive) COST Action community.[14] Additionally, it was distributed on scientific and professional social media platforms such as LinkedIn, ResearchGate, as well as other relevant contacts and mailing lists. The questionnaire is available as supplementary material accompanying this chapter via Zenodo (Amaro et al. 2026).

The survey remained open for one month, with a reminder sent to potential respondents after two weeks. We received 134 responses to the questionnaire.

[7] https://www.qualtrics.com/, last accessed in April 2025
[8] https://www.distant-reading.net/, last accessed in April 2025
[9] https://elex.is/, last accessed in April 2025
[10] https://www.elexicography.eu/, last accessed in April 2025
[11] https://euralex.org/, last accessed in April 2025
[12] https://www.cost.eu/actions/IC1302/, last accessed in April 2025
[13] https://typo.uni-konstanz.de/parseme/, last accessed in April 2025
[14] https://unidive.lisn.upsaclay.fr/, last accessed in April 2025

Most respondents belong to the academic sector (86%), with a smaller percentage from the corporate/private sector (10%) and the government/public sector (8%).[15]

The majority of respondents are experts in areas that typically bridge technology and language (see Figure 6).

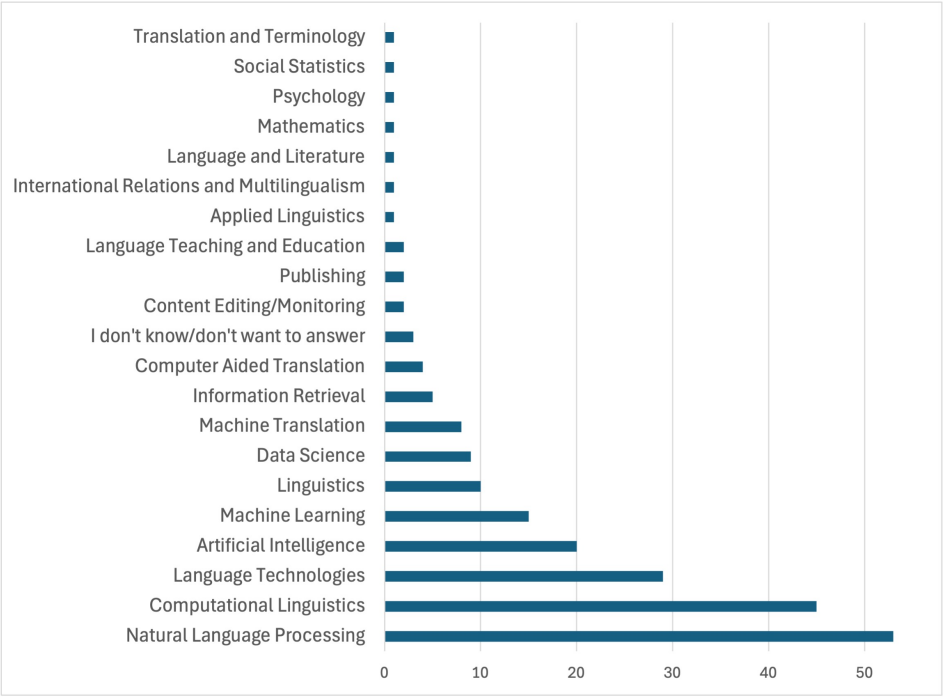

Figure 6: Users' areas of expertise

However, nearly a quarter of the respondents have other areas of expertise, such as linguistics – including syntax, semantics, lexicon, language acquisition, morphology, lexicography, terminology, and psycho-linguistics (10 respondents); language teaching and education (2 respondents); applied linguistics (1 respondent); international relations and multilingualism (1 respondent); language and literature (1 respondent); mathematics (1 respondent); psychology (1 respondent); social statistics (1 respondent); and translation and terminology (1 respondent). The majority of respondents are researchers or developers.

Out of the 134 respondents, only 79 answered the first MWE-specific question, namely whether they have worked with MWE lexicons or vocabularies for NLP tasks in the last 5 years. Of these, approximately 53% reported having worked

[15]Participants could select more than one sector of activity.

with MWE lexicons or vocabularies for NLP tasks in the past five years. Therefore, the data and results analysed in the following subsections primarily reflect the information provided by this subset of respondents.

6.2 Data and results

This subsection is organized according to the topics addressed by the survey questions, namely the use and applications of MWE lexicons, their performance and effectiveness in NLP tasks, insights and comments on improving and further using MWE lexicons, and information regarding the resources used by the respondents.

6.2.1 Use and applications of MWE lexicons

After the demographic questions concerning their professional background, respondents were asked if they had worked with MWE lexicons or vocabularies for NLP tasks in the past five years and, if so, to provide examples of the specific NLP tasks where these lexicons had been used. The responses revealed a rich variety of NLP tasks in which MWE lexicons have been applied over the past five years. Among the most frequently mentioned were parsing, followed closely by Machine Translation and Named Entity Recognition. Figure 7 presents the diversity and distribution of the NLP tasks reported. These tasks were applied in various contexts, as shown in Table 3.

Other tasks not necessarily related to NLP in which MWE lexicons were used include coreference resolution, idiomaticity representation in language models, metaphor extraction, quality monitoring, terminological checking, corpus revision, post-editing, encoding and standardisation, and language teaching/learning.

6.2.2 Evaluation of MWE lexicons

When asked to rate the effectiveness of the lexicons in enhancing the performance of the NLP tasks, most respondents did not provide an answer (42.9%). The remaining respondents rated the lexicons as moderately effective (26.2%), very effective (21.4%), extremely effective (7.1%), and not effective at all (2.4%).

The reasons for this evaluation may be linked to the specific features that the respondents considered relevant for the usefulness of MWE lexicons in NLP tasks, as well as the difficulties and challenges they encountered when using the lexicons. The generally useful characteristics of the lexicons mentioned include: being *freely accessible*, available in a *machine-usable format*, *multilingual*, having

Table 3: Contexts of applications of MWE lexicons.

Application	Link[a]
Annotation of verbal idioms	https://aclanthology.org/2024.mwe-1.16.pdf
Apertium Machine Translation	-
Collocations Dictionary of Modern Slovene	https://viri.cjvt.si/kolokacije/eng/
Computer-assisted language learning	-
Dictionary building	https://arxiv.org/abs/2005.10583
Electronic dictionaries and tools	http://www.institucional.grupogbd.com/-dicionario/index; http://www.dicesp.com:8083/
Entity extractor	https://string.hlt.inesc-id.pt/demo/entitiesextraction.pl
Lexicon building	https://doi.org/10.1017/S1351324921000401
Math term extraction	https://arxiv.org/abs/2307.06699, https://arxiv.org/abs/2208.13830, https://arxiv.org/abs/2311.12649, https://arxiv.org/abs/2309.0064
Metaphoric term extraction	http://journal.sepln.org/sepln/ojs/ojs/index.-php/pln/article/view/6558/3958
MWE-Finder	https://gretel5.hum.uu.nl/mwe-search
Named Entity Recogniser	https://www.juls.savba.sk/nerd/
Parser	https://portulanclarin.net/workbench/lx-udparser/
Semantic annotation for FrameNet	-
STRING–NLP Processing Chain	https://string.hlt.inesc-id.pt/
Testing a neural Machine Translation model	-
Text complexity assessment	https://iread4skills.com/

[a]All links were last accessed in March 2026.

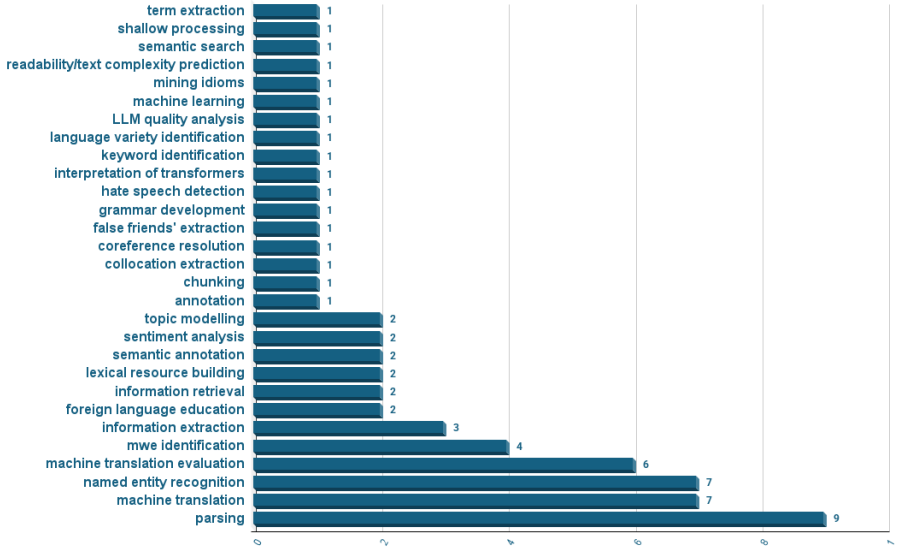

Figure 7: NLP tasks using MWE lexicons

wide coverage (both in terms of size and language), and having undergone *manual validation*. Regarding the information provided by these resources, the reported useful features include *canonical forms* and word forms, *morpho-syntactic* and *syntactic* information (such as Part-of-Speech (PoS), syntactic structure, lemma, or inflections), *semantic relations*, *sense/meaning* information, word *frequency*, *compositionality* information, *example* sentences, and *manual annotations* and comments.

The challenges and difficulties reported by respondents involved various issues, such as the *variation and instability* of resources over time, automatic *parsing errors*, *inconsistent definitions of MWEs* across resources, *non-standard* formats, inadequacies in automatically translated MWEs, *ambiguity*, and a *lack of usage examples* and specific information related to MWEs. This includes details such as which MWE elements undergo *inflection* and their *limitations*, the possibility and constraints of *inserting other words* within the MWE, and existing rules and *patterns*.

On the other hand, respondents noted several benefits of using MWE lexicons in NLP tasks, including:

- Increasing the coverage of resources and applications by accounting for MWEs that are not found in MWE-annotated corpora.

- Enhancing the reliability of MWE identification in corpora and improving the detection of both known and unknown MWEs.
- Improving parsing and various syntactic analyses, such as the analysis of verb constructions and verbal idioms.
- Enhancing coreference resolution by eliminating non-referential mentions.
- Enabling more accurate abstract semantic representations of texts and a deeper semantic understanding by capturing nuances and context.
- Improving the handling of MWEs by statistical models.
- Enhancing sentiment analysis.
- Improving readability assessment.
- Enhancing the automatic identification of idiomatic language.
- Assisting in obtaining specific information about the quality of neural machine translation.

6.2.3 Improving MWE lexicons and suggestions for further use

According to the respondents, existing resources could be improved to enhance their usefulness in NLP tasks in three key areas:

(i) **Coverage**: expanding MWE lexicons to include language varieties and dialects, specialized domains, and language registers (e.g., slangs), as well as implementing regular and dynamic updating processes.

(ii) **Information richness**: providing more detailed and finer-grained syntactic and semantic annotation of MWEs, along with adding more and better examples, including links to annotated corpora.

(iii) **Interoperability and access**: adopting a universal canonical form for MWEs and standardised representations, defining and adopting a universal MWE typology (e.g., as aimed for in the PARSEME project), ensuring licensing, broad distribution and compatible or convertible formats, and enabling cross-linguistic applications and search capabilities.

Finally, in response to the open suggestions and comments question, it was noted that MWE lexicons could be further utilized for domain adaptation, as part of lexicon customisation for specific domains, and to create benchmarks for testing LLMs on their ability to understand and process idiomaticity.

6.2.4 Details on MWE lexicons

In a dedicated section of the survey, respondents were asked to provide details on the lexical resources they used in NLP tasks in the past five years. Responses provided details on 24 electronic resources, ranging from encyclopedias to corpora. The units used to describe the resources varied significantly (e.g., MWE, lemma, collocation, file, MB, term, among others), which made it challenging to obtain comparable sizes. Among the four resources described using MWE as a unit, sizes ranged from 1,500 MWEs to 11,000 MWEs. The majority of the resources are monolingual (18 out of 24), while three are bilingual and another three are multilingual. Figure 8 below presents the languages covered by the described resources.

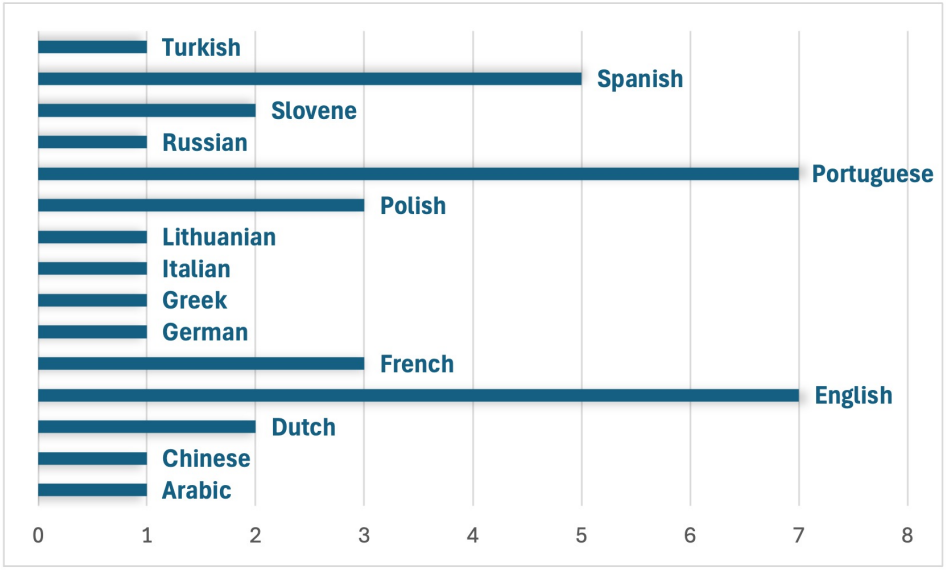

Figure 8: Languages accounted for in the described resources

All of the resources described are freely accessible for research purposes, with most being unrestricted or not governed by specific licensing terms. More information about these resources, including name and URL for access, is provided as supplementary material uploaded on Zenodo (Giouli et al. 2026b).

7 Discussion

The surveys conducted allowed us to gather information to address the four main research questions that set the framework for this chapter, briefly repeated here: (i) which tasks and applications MWE resources are used for; (ii) what types of information in the MWE resources are relevant for the respective tasks; (iii) what types of information MWE resources are lacking and should be included; and (iv) to what extent the use of these resources impacts the overall performance or tool. This section presents the analysis of the combined results from both surveys, highlighting areas of consensus and/or validation from users' responses and the literature, cases where the information is complementary, and instances of noticeable divergence.

7.1 Tasks and applications in which MWE resources are used

The NLP tasks in which MWEs were utilized largely overlap across the conducted surveys. Both surveys refer to tasks such as parsing, semantic annotation (e.g., sentiment labeling), MWE identification (including detection and extraction, but also modelling, recognition, and categorisation of MWEs), lexical resource building (through extraction or induction), grammar development (e.g., in the context of language generation), language variety identification (e.g., testing idiomaticity and improving language models to capture idiomaticity), foreign language education (e.g., systems for teaching and learning idioms and MWEs), machine translation (including neural machine translation and machine translation evaluation), LLM quality analysis (e.g., evaluation of LLM performance on MWEs), idiom mining (including idiomatic compound detection), and term extraction (including bilingual extraction). Additionally, results from the users' survey include the following tasks: shallow processing and chunking, coreference resolution, named entity recognition, false friends extraction, collocation extraction, topic modelling, keyword identification, information extraction and retrieval, semantic search, sentiment analysis, hate speech detection, readability and text complexity prediction, and machine learning and interpretation of transformers. The literature survey also included semantic relation extraction for terminology, and gender identification. The comprehensive list attests to the significance of MWEs and MWE resources in NLP tasks in general.

7.2 Relevant information in MWE resources

Both surveys report similar properties of the MWE resources in terms of (i) types of resources (including dictionaries, lexicons, wordnets, framenets, encyclopae-

dias, corpora, and various types of MWE datasets); (ii) number of languages covered by the resources (see Table 4); (iii) MWEs covered (nominal, verbal, adverbial, collocations, idioms, terms); and (iv) linguistic information provided in the resources (morphological, syntactic and semantic properties of the MWE components and/or the expressions; compositionality and context-dependency; frequency).

The broad features relevant to the use of MWEs in these tasks are accessibility, format, coverage, and manual validation. The contrast between the responses obtained from the users' survey and the information gathered in the literature review indicates that, in addition to basic and comprehensive information regarding MWE canonical forms and entries, and morpho-syntactic and syntactic information (PoS, syntactic structure, lemma, or inflectional variation), there is a focus on more sophisticated information. This includes linking to corpora for frequency and collocational data, examples in context, as well as information on compositionality, sense/meaning descriptions, and semantic relations.

Table 4: Coverage of languages in MWE resources

resource type	monolingual	bilingual	multilingual
users' survey	75%	12.5%	12.5%
literature survey	71.8%	12.8%	15.4%

7.3 Information to be included in MWE resources

An analysis of the two surveys shows that they are complementary in terms of the information to be included in MWE resources. Although this information can be directly related to the purposes and tasks for which the resources are used, there is a consensus that detailed and fine-grained syntactic and semantic annotations of MWEs should be provided. This includes information about the meaning and interpretation of MWEs, as well as their semantic classification of MWEs concerning compositionality or syntax/semantic classification of MWEs based on collocational properties.

Furthermore, responses from the users' survey highlight additional avenues for future developing, such as expanding MWE lexicons to encompass specialized domains, language varieties, and dialects, which can mean new information encoded in the resources. Suggestions also include linking MWEs to authentic/corpus data, through corpus annotation or concordances extraction, and creating or reusing existing resources to establish benchmarks for testing LLMs.

This could involve compiling specific corpora or utilizing declarative prompting to create a direct connection between MWEs and their meanings and usage restrictions.

All collected suggestions point to the need for and value of standards and references, whether related to MWE representations, whether in terms of universal/compatible typologies and formats, or whether concerning available training data. Interestingly, all suggestions appear to be language-independent and can be applied to various languages or even to multilingual contexts.

7.4 Assessing the impact of using MWE resources on the overall performance or tool

The users' survey was informative in what concerns the direct link between the resources used and their impact on the task at hand. Several cases and references were provided that illustrate the positive impact of MWE resources on parsing (e.g., Gonçalves et al. 2021, Müller et al. 2024), particularly concerning support-verb constructions (Baptista & Mamede 2020, Baptista et al. 2022), the treatment of the syntax of auxiliary verbs in Portuguese and Spanish (Baptista & Crismán 2021), and other verbal idioms (Galvão et al. 2019).

The use of MWE resources in the identification of MWEs in corpora or texts is also described as highly beneficial (Savary et al. 2019, Reis et al. 2021, Kroon & Odijk 2024). Other tasks in which the positive impact of MWE resources has been highlighted include language teaching (Eryiğit et al. 2022), particularly in applications such as controlling vocabulary associated with specific proficiency levels and readability assessment for educational placement (Da Corte & Baptista 2024).

As presented in Section 5, the literature survey addressed the impact of the use of MWE resources on specific tasks, with these being evaluated in the contexts of the studies reviewed. The tasks in which the positive impact of the use of MWE resources was reported included the identification of MWEs in corpora, but also translation and neural machine translation, vocabulary extraction, and language generation. Regarding the assessment of the use and respective impact of MWE resources, the surveys are complementary.

7.5 LLMs and the need for MWE lexicons

Although MWEs play an important role in NLP, their integration into pre-training and fine-tuning of LLMs is still limited. Approaches such as n-gram frequency analysis, BPE enhanced with Pointwise Mutual Information (Kumar

& Thawani 2022), and input compression by treating MWEs as single tokens (Gee et al. 2023) have been explored, but these methods are still experimental and primarily aim to improve efficiency. Multitoken prediction models (Gloeckle et al. 2024, DeepSeek-AI et al. 2025), which enable the prediction of multiple tokens at once or sequentially, are only used to a limited extent.

LLMs have been fine-tuned on annotated datasets to improve MWE classification, with models such as BERT, RoBERTa, DistilBERT, ALBERT and XLM-RoBERTa showing encouraging results. LLMs such as LLaMA 2, Mistral, and Phi-2 continue to have difficulties in distinguishing between literal and idiomatic meanings in zero-shot scenarios, especially in different languages (De Luca Fornaciari et al. 2024, De Leon et al. 2025). In contrast, smaller, fine-tuned models such as XLM-RoBERTa are still superior to these larger language models in MWE-related tasks.

Currently, the role of MWE lexicons can be seen in two ways: (i) for evaluating predictions from LLMs about whether a particular part of a sentence is complex and idiomatic; (ii) for annotating data with MWEs to fine-tune different types of LLMs. While their use with LLMs is still limited, the rapid advances in language modelling indicate a growing potential for the more effective incorporation of MWEs as they play a significant role in language complexity.

8 Conclusion

This survey underlines the importance of MWE lexicons in a wide range of NLP tasks. Based on a literature review and feedback from the community, it becomes clear that MWE resources are already being integrated into key tasks and applications such as syntactic parsing, machine translation, named entity recognition, and sentiment analysis - among others. And from this perspective, the users' perception survey aligns with previous surveys regarding the usefulness of MWE lexicons in different NLP tasks (Constant et al. 2017), while also extending their applicability to other tasks, such as named entity recognition, author gender identification, as well as understanding how LLMs work. In any case, the highest value of MWE lexicons resides in the linguistic richness they provide. These include morpho-syntactic and semantic annotations, usage examples, frequency data, and multilingual mappings. Lexicons that are manually validated and cover a wide range of languages are especially important. But despite that, evaluation of MWE lexicons in the context of NLP task implementation remains a challenge for the community, as evidenced by diverse user perceptions. This underscores the importance of developing standardised evaluation protocols and

methodological frameworks, including manual evaluation (Ramisch et al. 2023). Consistency and standardisation issues were also highlighted in the users' survey, particularly regarding the encoding of information, definitions, and annotation of MWEs, and other metadata types. Beyond the general call for enhancement of metadata descriptions (Losnegaard et al. 2016), users identified specific areas for improvement, including the coverage, richness of information, interoperability, and accessibility of MWE lexicons. These findings emphasize the need for continued advances in the development and evaluation of MWE lexicons to meet user and application demands more effectively. Finally, given that the papers analyzed primarily focused on improving the performance of specific NLP tasks, the impact of using dedicated and/or aware MWE resources was mentioned only when it positively affected the task. As expected, instances of negative or null impact were not reported. This selective reporting highlights the need for dedicated and comprehensive research on the broader impact of MWE integration in NLP tasks, which remains both necessary and highly relevant.

Abbreviations

ACL	Association for Computational Linguistics	NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers	PARSEME	Parsing and Multiword Expressions
BPE	Byte-Pair Encoding	PoS	Part of Speech
LLM	Large Language Model	STM	Structural Topic Modelling
MTP	Multitoken Prediction	VMWE	Verbal Multiword Expression
MWE	Multiword Expression		

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). The research of Raquel Amaro is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UID/3213/2025 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

References

Adjali, Omar, Emmanuel Morin & Pierre Zweigenbaum. N.d. Building comparable corpora for assessing multi-word term alignment. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3103–3112. Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.332/.

Amaro, Raquel, Verginica Barbu Mititelu, Voula Giouli, Svetla Koeva, Gražina Korvel, Irina Lobzhanidze & Giedre Valunaite Oleskeviciene. 2026. *Mwe lexicons in nlp tasks: a survey 2026 - questionnaire*. DOI: 10.5281/zenodo.21136030.

Avram, Andrei Marius, Verginica Barbu Mititelu & Dumitru-Clementin Cercel. 2023. Romanian multiword expression detection using multilingual adversarial training and lateral inhibition. In Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han & Shiva Taslimipoor (eds.), *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 7–13. Dubrovnik: Association for Computational Linguistics. DOI: 10.18653/v1/2023.mwe-1.4.

Baptista, Jorge. 2012. ViPEr: A lexicon-grammar of European Portuguese verbs. In Jan Radimsky (ed.), *Actes du 31e Colloque International sur le Lexique et la Grammaire* (Actes du 31e Colloque International sur le Lexique et la Grammaire), 10–16. Nové Hrady. https://hal.science/hal-01414265.

Baptista, Jorge & Rafael Crismán. 2021. Auxiliary verb constructions in Portuguese and Spanish. A comparative study. Construcciones verbales auxiliares en portugués y español. Un estudio comparativo. *Revista de Lenguas Modernas* 34. 39–57. DOI: 10.15517/rlm.v0i34.41462.

Baptista, Jorge & Nuno Mamede. 2020. Syntactic transformations in rule-based parsing of support verb constructions: Examples from European Portuguese. In Alberto Simões, Pedro Rangel Henriques & Ricardo Queirós (eds.), *9th Symposium on Languages, Applications and Technologies (SLATE 2020)*, 11:1–11:14. Leibniz: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Baptista, Jorge, Nuno Mamede & Sónia Reis. 2022. Support Verb Constructions across the Ocean Sea. In Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia & Carlos Ramisch (eds.), *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, 26–36. Marseille: European Language Resources Association. https://aclanthology.org/2022.mwe-1.6/.

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.

Bui, Van-Tuan & Agata Savary. 2024. Cross-type French multiword expression identification with pre-trained masked language models. 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)*, 4198–4204. Torino: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.374/.

Candito, Marie, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier & Silvio Ricardo Cordeiro. 2020. A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling* 8(2). 415–479. DOI: 10.15398/jlm.v8i2.265.

Capel, Annette. 2015. The English vocabulary profile. *English profile in practice* 5(1). 9–27.

Collard, Jacob, Valeria de Paiva & Eswaran Subrahmanian. 2023. *Parmesan: Mathematical concept extraction for education.* https://arxiv.org/abs/2307.06699.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.747. https://aclanthology.org/2020.acl-main.747/.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43(4). 837–892. DOI: 10.1162/COLI_a_00302.

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, 19–22. Marrakech: Association for Computational Linguistics.

Da Corte, Miguel & Jorge Baptista. 2024. Charting the linguistic landscape of developing writers: An annotation scheme for enhancing native language proficiency. 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*, 3046–3056. Torino: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.272.

de Jong, John, Mike Mayor & Catherine Hayes. 2016. Developing global scale of English learning objectives aligned to the Common European Framework. *Global Scale of English Research Series*.

De Leon, Frances Laureano, Harish Tayyar Madabushi & Mark G. Lee. 2025. *Evaluating large language models on multiword expressions in multilingual and code-switched contexts*. https://arxiv.org/abs/2504.20051.

De Luca Fornaciari, Francesca, Begoña Altuna, Itziar Gonzalez-Dios & Maite Melero. 2024. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty & Emmy Liu (eds.), *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, 35–44. Mexico City: Association for Computational Linguistics. DOI: 10.18653/v1/2024.figlang-1.5. https://aclanthology.org/2024.figlang-1.5/.

DeepSeek-AI et al. 2025. *DeepSeek-v3 Technical Report*. https://arxiv.org/abs/2412.19437.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran & Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 4171–4186. Association for Computational Linguistics. DOI: 10.18653/V1/N19-1423.

Eryiğit, GülŞen, Ali Şentaş & Johanna Monti. 2022. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering* 29. 1–33. DOI: 10.1017/S1351324921000401.

Espinosa Anke, Luis, Steven Schockaert & Leo Wanner. 2019. Collocation classification with unsupervised relation vectors. In Anna Korhonen, David Traum & Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5765–5772. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/P19-1576.

Espinosa-Anke, Luis, Joan Codina-Filba & Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association*

for Computational Linguistics: Main Volume, 1406–1417. Association for Computational Linguistics.

Fellbaum, Christiane. 2010. WordNet. In Roberto Poli, Michael Healy & Achilles Kameas (eds.), *Theory and applications of ontology: Computer applications*, 231–243. Dordrecht: Springer Netherlands. DOI: 10.1007/978-90-481-8847-5_10.

Fisas, Beatriz, Luis Espinosa Anke, Joan Codina-Filbá & Leo Wanner. 2020. CollFrEn: Rich bilingual English–French collocation resource. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 1–12. Association for Computational Linguistics.

Gage, Philip. 1994. A new algorithm for data compression. *The C Users Journal* 12(2). 23–38.

Galvão, Ana, Jorge Baptista & Nuno Mamede. 2019. New developments on processing European Portuguese verbal idioms. In Carlos Augusto Prolo (ed.), *STIL-Symposium in Information and Human Language Technology*, 229–238. Brazil: Brazilian Computer Society, Brazilian Special Interest Group on Natural Language Processing.

Garcia, Marcos, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In Paola Merlo, Jorg Tiedemann & Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 3551–3564. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.310.

Gee, Leonidas, Leonardo Rigutini, Marco Ernandes & Andrea Zugarini. 2023. Multi-word tokenization for sequence compression. In Mingxuan Wang & Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry track*, 612–621. Singapore: Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-industry.58.

Giouli, Voula, Verginica Barbu Mititelu, Raquel Amaro, Svetla Koeva, Gražina Korvel, Irina Lobzhanidze & Giedre Valunaite Oleskeviciene. 2026a. *MWE lexicons in NLP tasks: a survey 2026 - Lexical Resources in Papers*. Zenodo. DOI: 10.5281/zenodo.21135928.

Giouli, Voula, Verginica Barbu Mititelu, Raquel Amaro, Svetla Koeva, Gražina Korvel, Irina Lobzhanidze & Giedre Valunaite Oleskeviciene. 2026b. *MWE lexicons in NLP tasks: a survey 2026 - Lexical Resources in Questionnaire*. Zenodo. DOI: 10.5281/zenodo.21136185.

Giouli, Voula, Verginica Barbu Mititelu, Raquel Amaro, Svetla Koeva, Gražina Korvel, Irina Lobzhanidze & Giedre Valunaite Oleskeviciene. 2026c. *MWE lexicons in NLP tasks: a survey 2026 - Paper Analysis*. Zenodo. DOI: 10.5281/zenodo.21108051.

Gloeckle, Fabian, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz & Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett & Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning*. Vienna: JMLR.org. https://openreview.net/forum?id=pEWAcejiU2.

Gonçalves, Matilde, Luisa Coheur, Jorge Baptista & Ana Mineiro. 2021. Avaliação de recursos computacionais para o português. *Linguamática* 12(2). 51–68. DOI: 10.21814/lm.12.2.331.

Haagsma, Hessel, Malvina Nissim & Johan Bos. 2018. The other side of the coin: Unsupervised disambiguation of potentially idiomatic expressions by contrasting senses. In Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan & Miriam R. L. Petruck (eds.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 178–184. Santa Fe: Association for Computational Linguistics. https://aclanthology.org/W18-4919/.

Haviv, Adi, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg & Mor Geva. N.d. Understanding transformer memorization recall through idioms. In Andreas Vlachos & Isabelle Augenstein (eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics*, 248–264. Dubrovnik: Association for Computational Linguistics. DOI: 10.18653/v1/2023.eacl-main.19. https://aclanthology.org/2023.eacl-main.19/.

Henrich, Verena & Erhard Hinrichs. 2011. Determining immediate constituents of compounds in GermaNet. In Ruslan Mitkov & Galia Angelova (eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 420–426. Hissar: Association for Computational Linguistics. https://aclanthology.org/R11-1058/.

Hubers, Ferdy, Catia Cucchiarini & 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*, 4418–4425.

Hwang, Alyssa & Christopher Hidey. 2019. Confirming the non-compositionality of idioms for sentiment analysis. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 125–129.

Jakhotiya, Yash, Vaibhav Kumar, Ashwin Pathak & Raj Shah. N.d. JARVix at SemEval-2022 Task 2: It takes one to know one? Idiomaticity detection using zero and one-shot learning. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh & Shyam Ratan (eds.), *Proceedings of the 16th International Workshop on Seman-*

tic Evaluation (SemEval-2022), 165–168. Seattle: Association for Computational Linguistics. DOI: 10.18653/v1/2022.semeval-1.19.

Jhamtani, Harsh, Varun Gangal, Eduard Hovy & Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wentau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7476–7485. Punta Cana: Association for Computational Linguistics. DOI: 10.18653/v1/2021.emnlp-main.592.

Jin, Zhijing, Yuen Chen, Fernando Gonzalez Adauto, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf & Mona Diab. 2024. Analyzing the role of semantic representations in the era of large language models. In Kevin Duh, Helena Gomez & Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3781–3798. Mexico City: Association for Computational Linguistics. DOI: 10.18653/v1/2024.naacl-long.209.

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/.

Kanclerz, Kamil & Maciej Piasecki. 2022. Deep neural representations for multiword expressions detection. In Samuel Louvan, Andrea Madotto & Brielen Madureira (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 444–453. Dublin: Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-srw.36.

Knietaite, Agne, Adam Allsebrook, Anton Minkov, Adam Tomaszewski, Norbert Slinko, Richard Johnson, Thomas Pickard, Dylan Phelps & Aline Villavicencio. 2024. Is less more? Quality, quantity and context in idiom processing with natural language models. *arXiv preprint arXiv:2405.08497*.

Kokkinakis, Dimitrios, Ricardo Muñoz Sánchez, Sebastianus Bruinsma & Mia-Marie Hammarlin. 2023. Investigating the effects of MWE identification in structural topic modelling. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 36–44.

Kroon, Martin & Jan Odijk. 2024. MWE-Finder: An evaluation through three case studies. In Krister Lindén, Thalassia Kontino & Jyrki Niemi (eds.), *Selected papers from the CLARIN Annual Conference 2023*. Linköping Electronic Conference Proceedings 210. DOI: 10.3384/ecp210009.

Krotova, Irina, Sergey Aksenov & Ekaterina Artemova. 2020. A joint approach to compound splitting and idiomatic compound detection. 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*, 4410–4417. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.543/.

Kumar, Dipesh & Avijit Thawani. 2022. BPE beyond word boundary: How NOT to use multi word expressions in neural Machine translation. In Shabnam Tafreshi, João Sedoc, Anna Rogers, Aleksandr Drozd, Anna Rumshisky & Arjun Akula (eds.), *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, 172–179. Dublin: Association for Computational Linguistics. DOI: 10.18653/v1/2022.insights-1.24. https://aclanthology.org/2022.insights-1.24.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma & Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020*. Addis Ababa. https://openreview.net/forum?id=H1eA7AEtvS.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier & Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. 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*, 2479–2490. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.302/.

Lee, John S. Y. & Adilet Uvaliyev. 2023. Automatic generation of vocabulary lists with multiword expressions. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 81–86.

Liu, Changsheng & Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In Kevin Knight, Ani Nenkova & Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Com-*

putational Linguistics: Human Language Technologies, 363–373. San Diego: Association for Computational Linguistics. DOI: 10.18653/v1/N16-1040.

Liu, Li & Francois Lareau. 2024. Assessing BERT‘s sensitivity to idiomaticity. 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*, 14–23. Torino: ELRA & ICCL. https://aclanthology.org/2024.mwe-1.4/.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692. http://arxiv.org/abs/1907.11692.

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/.

Lu, Daming. 2022. Daminglu123 at SemEval-2022 task 2: Using BERT and LSTM to do text classification. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh & Shyam Ratan (eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 186–189. Seattle: Association for Computational Linguistics. DOI: 10.18653/v1/2022.semeval-1.22.

Mahajan, Ujwala P., Ajay S. Patil & Nita V. Patil. 2024. A survey of tools and techniques for multiword expression detection. *International Journal of Computer Applications* 186(32). 11–18. DOI: 10.5120/ijca2024923851.

Markantonatou, Stella, 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.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah & Benoît Sagot. 2020. CamemBERT: A tasty French language model. In Dan Jurafsky, Joyce Chai, Natalie

Schluter & Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.645.

Maziarz, Marek, Ewa Rudnicka & Łukasz Grabowski. 2022. Multi-word lexical units recognition in wordnet. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC 2022*, 49–54. Association for Computational Linguistics.

Mel'čuk, Igor A. 2007. 11. Lexical functions. In Harald Burger, Dmitrij Dobrovolskij, Peter Kühn & Neal R. Norrick (eds.), *Phraseologie*, vol. 1, 119–131. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110197136-014.

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. https://aclanthology.org/2024.tacl-1.33/.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. 1990. Introduction to WordNet: An online lexical database. *International Journal of Lexicography* 3(4). 235–244.

Mohamed, Najet Hadj, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar Keskes & Lamia Hadrich Belguith. 2024. Lexicons gain the upper hand in Arabic MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024*, 88–97.

Moreno-Ortiz, Antonio, Javier Fernandez-Cruz & Chantal Pérez Chantal Hernández. 2020. Design and evaluation of SentiEcon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. 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*, 5065–5072. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.623/.

Müller, Izabela, Nuno Mamede & Jorge Baptista. 2022. Bootstrapping a lexicon of multiword adverbs for Brazilian Portuguese. In Gloria Corpas Pastor & Ruslan Mitkov (eds.), *Computational and corpus-based phraseology*, 160–174. Cham: Springer International Publishing.

Müller, Izabela, Nuno Mamede & Jorge Baptista. 2024. Hurdles in parsing multiword adverbs: Examples from Portuguese. In Pablo Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira & Raquel Amaro (eds.), *Proceedings of the 16th International Conference on Computational*

Processing of Portuguese -, vol. 1, 177–186. Santiago de Compostela: Association for Computational Linguistics. https://aclanthology.org/2024.propor-1.18.

Negi, Gaurav, Rajdeep Sarkar, Omnia Zayed & Paul Buitelaar. N.d. A hybrid approach to aspect based sentiment analysis using transfer learning. 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)*, 647–658. Torino: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.56/.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang & Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha: Association for Computational Linguistics. DOI: 10.3115/v1/D14-1162.

Pershina, Maria, Yifan He & Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In Michael Roth, Annie Louis, Bonnie Webber & Tim Baldwin (eds.), *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, 76–82. Lisbon: Association for Computational Linguistics. DOI: 10.18653/v1/W15-2709.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn Walker, Heng Ji & Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, vol. 1, 2227–2237. New Orleans: Association for Computational Linguistics. DOI: 10.18653/v1/N18-1202. https://aclanthology.org/N18-1202/.

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.

Premasiri, Damith, Amal Haddad Haddad, Tharindu Ranasinghe & Ruslan Mitkov. 2022. Transformer-based detection of multiword expressions in flower and plant names. *arXiv preprint arXiv:2209.08016*.

Premasiri, Damith, Amal Haddad Haddad, Tharindu Ranasinghe & Ruslan Mitkov. 2023. Deep learning methods for identification of multiword flower and plant names. In Ruslan Mitkov & Galia Angelova (eds.), *Proceedings of the*

14th International Conference on Recent Advances in Natural Language Processing, 879–887. Varna: INCOMA Ltd. https://aclanthology.org/2023.ranlp-1.95/.

Prochnow, Alexander, Johannes E. Bendler, Caroline Lange, Foivos Ioannis Tzavellos, Bas Marco Göritzer, Marijn ten Thij & Riza Batista-Navarro. 2024. IDEM: The IDioms with EMotions dataset for emotion recognition. 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)*, 8569–8579. Torino: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.752/.

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya & Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan & Miriam R. L. Petruck (eds.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 222–240. Santa Fe: Association for Computational Linguistics. https://aclanthology.org/W18-4925.

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/.

Ramisch, Carlos, Abigail Walsh, Thomas Blanchard & Shiva Taslimipoor. 2023. A survey of MWE identification experiments: The devil is in the details. In Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han & Shiva Taslimipoor (eds.), *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 106–120. Dubrovnik: Association for Computational Linguistics. DOI: 10.18653/v1/2023.mwe-1.15.

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.

Reis, Sónia, Jorge Baptista & Nuno J. Mamede. 2021. Provérbios portugueses usuais: Distribuição em corpora. In Evandro Eduardo Seron Ruiz & Tiago Timponi Torrent (eds.), *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 325–334. SBC.

Rosen, Victoria, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejcek, Agata Savary, Adam Przepiorkowski, Petya Osenova & Verginica Barbu Mititelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the treebanks and linguistic theories conference (TLT - 2015)*, 179–193. Warsaw. https : / / ufal . mff . cuni . cz / biblio / attachments / 2015 - bejcek - m9168853603341436530.pdf.

Rudnicka, Ewa Katarzyna, Wojciech Witkowski & Michał Kaliński. 2015. Towards the methodology for extending princeton WordNet. *Cogn. Stud.* 15. 335–351.

Sanh, Victor, Lysandre Debut, Julien Chaumond & Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108. http://arxiv.org/abs/1910.01108.

Sarwar, Raheem. 2022. Author gender identification for urdu articles. In Gloria Corpas Pastor & Ruslan Mitkov (eds.), *International Conference on Computational and Corpus-Based Phraseology*, 221–235. Malaga: Springer.

Savary, Agata, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics* 112. 5–54. DOI: 10.2478/pralin-2019-0001.

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.

Schick, Timo & Hinrich Schütze. 2020. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 3996–4007. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.368.

Schick, Timo & Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jorg Tiedemann & Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.20.

Schneider, Nathan, Dirk Hovy, Anders Johannsen & Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov & Torsten Zesch (eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 546–559. San Diego: Association for Computational Linguistics. DOI: 10.18653/v1/S16-1084.

Schulte im Walde, Sabine. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources*, 269–308. Berlin: Language Science Press.

Schulte im Walde, Sabine, Anna Hätty, Stefan Bott & Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compounds. 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)*, 2285–2292. Portorož: European Language Resources Association (ELRA). https://aclanthology.org/L16-1362.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (: Long papers)*, vol. 1, 1715–1725. Berlin: Association for Computational Linguistics. DOI: 10.18653/v1/P16-1162.

Tanner, Joshua & Jacob Hoffman. 2023. MWE as WSD: Solving multiword expression identification with word sense disambiguation. In Houda Bouamor, Juan Pino & Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, 181–193. Singapore: Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-emnlp.14. https://aclanthology.org/2023.findings-emnlp.14/.

Tayyar Madabushi, Harish, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh & Shyam Ratan (eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 107–121. Seattle: Association for Computational Linguistics. DOI: 10.18653/v1/2022.semeval-1.13.

Tayyar Madabushi, Harish, Edward Gow-Smith, Carolina Scarton & Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and Methods for the Exploration of idiomaticity in Pre-Trained Language Models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wen-tau Yih (eds.), *Findings of the association for computational linguistics: EMNLP 2021*, 3464–3477. Punta Cana: Association for Computational Linguistics. DOI: 10.18653/v1/2021.findings-emnlp.294. https://aclanthology.org/2021.findings-emnlp.294/.

Yang, Jinbiao. 2024. *Rethinking tokenization: Crafting better tokenizers for large language models*. https://arxiv.org/abs/2403.00417.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov & Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Edward A. Fox & Roman Garnett (eds.), *Advances in neural information processing systems*, vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Zhou, Jianing, Hongyu Gong & Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, 33–48.

Supplementary Material

To comply with the publisher's length restrictions, the extensive supporting material accompanying this chapter has been deposited in Zenodo. The supplementary resources contain the complete paper analysis, inventories of lexical resources, and the survey questionnaire referred to throughout the chapter.

MWE Lexicons in NLP Tasks: A Survey (2026) – Paper Analysis. DOI: https://doi.org/10.5281/zenodo.21108051

MWE Lexicons in NLP Tasks: A Survey (2026) – Lexical Resources in Papers. DOI: https://doi.org/10.5281/zenodo.21135928

MWE Lexicons in NLP Tasks: A Survey (2026) – Questionnaire. DOI: https://doi.org/10.5281/zenodo.21136030

MWE Lexicons in NLP Tasks: A Survey (2026) – Lexical Resources in Questionnaire. DOI: https://doi.org/10.5281/zenodo.21136185

All these links to resources were last accessed in April 2025.