PARSEME 2.0 Multilingual Corpus of Multiword Expressions

Agata Savary[1], Manon Scholivet[1], Carlos Ramisch[2], Takuya Nakamura[1,3], Eric Bilinski[1], Sara Stymne[4], Voula Giouli[5], Stella Markantonatou[6], Vasile Păiș[7], Maria Mitrofan[7], Louis Estève[1], Bruno Guillaume[8], Verginica Barbu Mititelu[7], Jaka Čibej[9], Roberto A. Díaz Hernández[10], Victoria Fendel[11], Polona Gantar[9], Olha Kanishcheva[12], Cvetana Krstev[13], Chaya Liebeskind[14], Irina Lobzhanidze[15], Aleksandra Marković[16], Gunta Nešpore-Bērzkalne[17], Adriana Pagano[18], Mehrnoush Shamsfard[19], Ranka Stanković[20], Vahide Tajalli[19], Carole Tiberius[21], Aakanksha Padhye[22]

[1]LISN, Paris-Saclay University, CNRS, France; [2]Aix Marseille Univ, CNRS, LIS, Marseille, France;
[3]LIGM, Université Gustave-Eiffel, France; [4]Uppsala University, Sweden;
[5]Aristotle University of Thessaloniki and ILSP, ATHENA RC, Greece;
[6]ILSP and Archimedes Unit, ATHENA RC, Greece; [7]RACAI, Bucharest, Romania;
[8]Université de Lorraine, CNRS, Inria, LORIA, France; [9]University of Ljubljana, Slovenia;
[10]University of Jaen, Spain; [11]University of Oxford, United Kingdom;
[12]Heidelberg University, Germany; SET University, Ukraine;
[13]Jerteh - Language Resources and Technologies Society, Serbia;
[14]Jerusalem College of Technology, Israel; [15]Ilia State University, Tbilisi, Georgia;
[16]Institute for the Serbian Language SASA, Belgrade, Serbia;
[17]Institute of Mathematics and Computer Science, University of Latvia;
[18]Federal University of Minas Gerais, Brazil; [19]Shahid Beheshti University, Tehran, Iran;
[20]University of Belgrade, Serbia;
[21]Dutch Language Institute/Leiden University Centre for Linguistics, The Netherlands;
[22]Indian Institute of Technology Delhi, India

Abstract

We present edition 2.0 of the PARSEME multilingual corpus annotated for multiword expressions (MWEs), resulting from efforts of the PARSEME community towards universality-driven modeling of idiomaticity. With respect to previous editions, we extend the annotation scope to all syntactic MWE categories: verbal, nominal, adjectival, adverbial and functional. We cover 17 languages, of which 7 are new. The annotation process is based on cross-lingually unified guidelines, phrased as decision diagrams over linguistic tests, and a typology of 18 MWE categories. The corpus contains almost 5 million tokens, over 250,000 sentences and 140,000 MWE annotations. The applicability of the corpus is tested in baseline experiments with a prompt-based MWE identification system. Results show that generic large language models do not encode sufficient knowledge to solve the MWE identification task.

Keywords: multiword expressions, PARSEME, multilingual corpus

1. Introduction

Multiword expressions (MWEs), such as ***by and large***, or ***cut corners***,[1] defy regular rules of language and have been considered challenging in natural language processing (NLP) for decades (Constant et al., 2017). Their most outstanding property is semantic non-compositionality, i.e. the fact that their meaning cannot be inferred straightforwardly from the meanings of their components and from their syntactic structure, e.g. *to* ***cut corners*** 'to do sth poorly in order to save costs' has few explicit links with literally cutting corners.

The rise of distributional semantics brought new opportunities for modeling non-compositionality of MWEs (Salehi et al., 2015; Schulte im Walde et al., 2016; Cordeiro et al., 2019). However, despite overall growing capacities of transformers and other language models, MWEs remain difficult to represent and process. They resist generalization beyond seen data (Savary et al., 2019). Their semantics is not properly captured by language models (Shwartz and Dagan, 2019), which instead rely on surface patterns and memorization (Miletić and Schulte im Walde, 2024), and require significant computational resources for fine-tuning to come close to human performance (Pickard et al., 2025). Moreover, relatively few efforts were dedicated to understanding MWEs in multilingual settings. The main obstacle is the scarceness of MWE-aware language resources which would make cross-language comparisons possible.

We describe a resource that bridges this gap: the PARSEME 2.0 multilingual corpus annotated for MWEs, distributed under open licenses. It

[1]MWE examples follow Markantonatou et al. (2021).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4819–4834
11–16 May 2026.

is the culmination of long-standing efforts of the PARSEME community towards cross-lingually unified linguistic modeling of idiomaticity. The 4 previous releases of the corpus were dedicated to verbal MWEs only. The current edition, for the first time, covers MWEs of all syntactic categories: verbal (***cut corners***), nominal (***hot dog***), adjectival (***clear-cut***), adverbial (***by and large***), and functional (***provided that***). We cover 17 languages, including 10 (out of 26) covered previously (called *veteran* languages) – French (fr), Modern Greek (el), Hebrew (he), Persian (fa), Polish (pl), Portuguese (pt), Romanian (ro), Slovene (sl), Swedish (sv), Serbian (sr) – and 7 new ones – Dutch (nl), Egyptian (egy), Georgian (ka), Ancient Greek (grc), Japanese (ja), Latvian (lv) and Ukrainian (uk). One language, Marathi (mr), is still in the release pipeline and will be published soon. All datasets have been manually annotated based on cross-lingually unified guidelines, and build upon morpho-syntactic annotations compatible with Universal Dependencies (Nivre et al., 2020). The corpus is available under open licenses.[2]

2. State of the art

The special role of MWEs in natural language has been acknowledged for decades notably via efforts put into language resources such as MWE lexicons and MWE-aware treebanks (Losnegaard et al., 2016; Rosén et al., 2016). Most of them were unilingual and addressed MWEs to a limited degree, as part of more general modeling issues.

With growing interest in MWE-dedicated tasks in NLP, several corpora focusing explicitly on MWE annotation appeared. They were usually monolingual and limited to selected categories, e.g. English idiomatic verb-noun combinations (Fazly and Stevenson, 2006; Cook et al., 2007), Estonian particle verbs (Kaalep and Muischnek, 2008), French nominal and adverbial MWEs (Laporte et al., 2008b,a), Hungarian light-verb constructions (Vincze and Csirik, 2010), Arabic verb-noun and verb-particle constructions (Bar et al., 2014).

PARSEME in its 4 previous editions, 1.0, 1.1, 1.2 and 1.3 (Savary et al., 2017; Ramisch et al., 2018, 2020; Savary et al., 2023) was larger than these previous efforts both in terms of multilingualism (18, 19, 14, and 26 languages were covered, respectively) and of the MWE categories, since it accounted for all types of verbal MWEs. However, non-verbal categories were still not considered.

Datasets with all MWEs categories are available for some languages. The English DiMSUM corpus (Schneider et al., 2016) contains 5,069 verbal and non-verbal MWEs annotated following (Schneider et al., 2014). The Turkish MWE-annotated corpus is inspired from DiMSUM (Adalı et al., 2016). The French Sequoia corpus is annotated for MWEs and named entities, with particular care taken in drafting the annotation guidelines, organized in flowcharts, in the spirit of PARSEME 1.1 and 1.2, but extended to non-verbal MWEs (Candito et al., 2021). The Sequoia guidelines were, a major inspiration for the PARSEME 2.0 guidelines (Sec. 4). The more recent English CoAM corpus (Ide et al., 2025) contains 1,3K sentences, annotated for the MWE categories from Schneider et al. (2014) with semi-automatic methods.

Other MWE datasets covering a few languages were created for shared tasks related to disambiguation of literal and idiomatic uses of MWEs in English and German (Korkontzelos et al., 2013), semantic text similarity in English, Portuguese and Galician (Tayyar Madabushi et al., 2022), and MWE comprehension in multimodal settings in English and Portuguese (Pickard et al., 2025).

ID10M (Tedeschi et al., 2022) is a dataset in 10 languages, built from Wiktionary MWEs entries of various categories. Occurrences are extracted automatically from the examples appearing in these entries, and from a large raw corpus. In 4 languages, annotations are manually curated. ID10M is limited to MWEs from Wiktionary (where no unified definition of a MWE is shared across contributors), it does not contain any information on MWE categories, nor does it include MWEs with gaps. Despite these shortcomings, it remains useful for tasks such as MWE identification. In Sec. 9 we compare this work to ours in terms of diversity.

It the light of this state of the art, it appears that a multilingual, cross-lingually comparable corpus annotated for MWEs of all categories is missing so far. The next sections describe the efforts of the PARSEME community to bridge this gap.

3. People

In PARSEME, the construction of the corpus and community building continuously enrich each other. Each language is represented by one or more Language Leaders (LLs), most often affiliated in the countries where their respective languages are spoken. LLs recruit and coordinate annotators, select and prepare texts to annotate, and check consistency of annotations. All annotators are native speakers, with three exceptions: non-spoken languages (Ancient Greek and Old Egyptian) and French (where a near-native expert is part of the team). The PARSEME Core Group implements and maintains centralized infrastructure (Gitlab repositories, online guidelines, annotation platform, format validators and converters, tools for consistency checks, corpus split and release,

[2] http://hdl.handle.net/11372/LRT-6123.

<table>
<tr><td rowspan="8">VMWE</td><td>VID</td><td>verbal idiom</td><td>(ka) შიშს ჭამს (šišs čams | lit. ‘he eats horror’) ‘he panics’</td></tr>
<tr><td>LVC.full</td><td>light-verb construction; bleached verb</td><td>(nl) een toespraak houden (lit. ‘to hold a speech’) ‘to make a speech’</td></tr>
<tr><td>LVC.cause</td><td>light-verb construction; causal verb</td><td>(sl) narediti konec nečemu (lit. ‘to make an end to sth’) ‘to end sth’</td></tr>
<tr><td>IRV</td><td>inherently reflexive verb</td><td>(sr) бојати се (bojati se | lit. ‘fear oneself’) ‘to be afraid’</td></tr>
<tr><td>IVPC.full</td><td>idiomatic verb-particle construction</td><td>(el) βάζω μπρος (vazo bros | lit. ‘to put forward’) ‘to start’</td></tr>
<tr><td>IVPC.semi</td><td>semi-idiomatic verb-particle constr.</td><td>(sv) fråga ut (lit. ‘to ask out’) ‘to interrogate / to invite out’</td></tr>
<tr><td>MVC</td><td>multi-verb construction</td><td>(egy) (šm.t iw.t nsw | lit. ‘going coming king’) ‘king paying a visit’</td></tr>
<tr><td>IAV</td><td>inherently adpositional verb</td><td>(uk) вилився в те (vylyvsya v te | lit. ‘spilled into this’) ‘resulted in this’</td></tr>
<tr><td rowspan="3">NMWE</td><td>NID</td><td>nominal idiom</td><td>(mr) हातपाय (hātapāya | lit. ‘hand-feet’) ‘limbs’</td></tr>
<tr><td>PronID</td><td>pronominal idiom</td><td>(ja) 何でも (lit. ‘what be even’) ‘whatever’</td></tr>
<tr><td>NV</td><td>deverbal nominal MWE</td><td>(lv) kāju atstiepšana (lit. ‘stretching of one’s legs’) ‘dying’</td></tr>
<tr><td rowspan="3">AMWE</td><td>AdjID</td><td>adjectival idiom</td><td>(grc) καλοὶ κἀγαθοὶ (lit. ‘good and beautiful’) ‘physically and morally excellent’</td></tr>
<tr><td>AdvID</td><td>adverbial idiom</td><td>(pl) zrobić coś raz dwa (lit. ‘to do sth one two’) ‘to do sth quickly’</td></tr>
<tr><td>AV</td><td>deverbal adjectival/adverbial MWE</td><td>(fa) دل‌نشین (del neshin | lit. ‘sitting on the heart’) ‘pleasant’</td></tr>
<tr><td rowspan="4">FuncMWE</td><td>DetID</td><td>determiner idiom</td><td>(ro) tot felul de demersuri (lit. ‘all kinds of steps’) ‘various steps’</td></tr>
<tr><td>AdpID</td><td>adpositional idiom</td><td>(fr) hors de danger ‘out of danger’</td></tr>
<tr><td>ConjID</td><td>conjunction idiom</td><td>(he) כמו כן (kmo ken | lit. ‘as so’) ‘likewise’</td></tr>
<tr><td>IntjID</td><td>interjection idiom</td><td>(pt) aí está (lit. ‘there is’) ‘here you are’</td></tr>
</table>

Table 1: PARSEME typology of MWEs with multilingual examples.

documentation, etc.).

Weekly online meetings bring conviviality and momentum to the process. Yearly meetings in person are possible due to funding from COST[3]. These instruments enable us to progressively build a linguistically and culturally diverse, universality-driven network, where core linguistic and NLP expertise is enrooted in the native research communities, thus relatively sustainable.

Specifically behind edition 2.0 of the corpus, the team is composed of 21 LLs, 48 annotators[4] (see App. 15.2) and 5 core group experts. Across all editions, this group has over 240 members.

4. Annotation guidelines

Previous versions of the PARSEME corpora (1.0–1.3) only cover verbal MWEs. Version 2.0 is a considerable extension, covering all MWE categories, also including nominal, adjectival, adverbial, and functional MWEs.[5] The full tagset now includes 18 categories illustrated in Tab. 1. The first version of guidelines 2.0 was authored by a small group of experts, who started from version 1.3 for verbal MWEs. This was followed by a pilot annotation for a range of languages, discussions upon arising issues, enhancements of the guidelines and adding language-specific examples to the new tests.

A major principle in PARSEME, inspired by linguistic studies (Gross, 1986, 1988; Nunberg et al., 1994), is that semantic non-compositionality is hard to test directly but correlates with lexical and morpho-syntactic inflexibility. For instance, in (fr) ***à la place de*** *Luc* (lit. ‘at the place of Luc’) ‘instead of Luc’, the expression in bold does not allow *place* to be in plural (*aux places de Luc* ‘at the places of Luc’ can only be interpreted literally), although noun inflection for number is a regular process in French. This morphological inflexibility gives this expression the status of a MWE. Linguistic tests like this one, illustrated by examples in various languages, as shown in Fig. 1, are atomic components of the guidelines. Most tests are language-independent, but a few are language-specific, e.g., for inherently clitic verbs, like (it) ***prender|le*** (lit. ‘take them’) ‘get beaten up’, or multiword tokens, like (sv) ***av|bryta*** (lit. ‘off-break’) ‘cancel’.

Another principle in PARSEME is reproducibility of the annotation process. Guidelines are organized into decision diagrams (also called *branches*), so that, for two annotators examining the same MWE candidate, if their answers to the tests are the same, the outcome of the annotation is also the same. Fig. 2 shows the decision diagram dedicated to functional MWEs, where *if-then-else* constructions upon tests, like the one in Fig. 1, redirect either to a decision about a MWE category or to other tests. Similar diagrams exist for adjectival-adverbial, verbal and nominal MWE candidates. The verbal diagram is the one from edition 1.3.[6] The nominal one includes tests, strongly inspired by Candito et al. (2021), discarding named entities, which are beyond our scope.

The decision diagrams are applied manually by annotators on running text. They first identify an expression that could potentially be an MWE, transform it into a so-called neutral form, then go to a unique entry point, where the distribution of this

[3]European Cooperation in Science and Technology (https://www.cost.eu). PARSEME originates from the homonymous COST IC1207 action (2013-2017). The same community is now supported by COST CA21167 action UniDive (https://unidive.lisn.upsaclay.fr/)

[4]Most LLs are also annotators.

[5]https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

[6]Up to renaming VPC to IVPC.

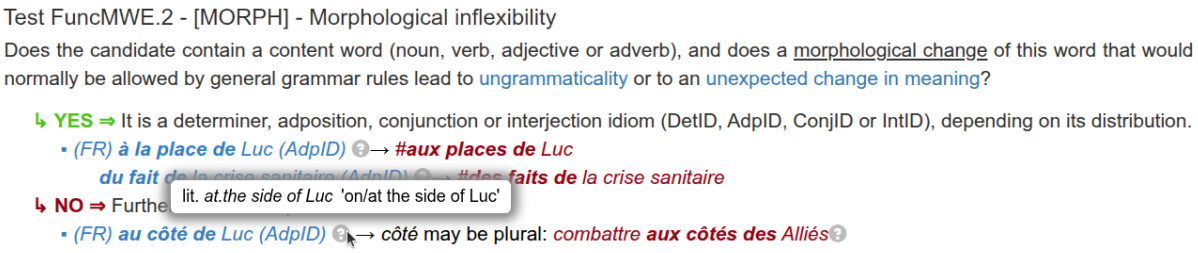

Figure 1: Morphological inflexibility test for functional MWEs, with examples in French.

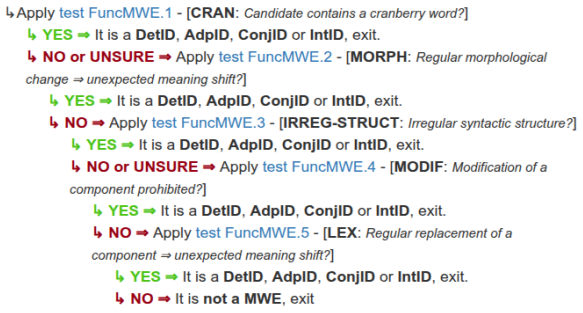

Figure 2: Decision diagram for functional MWEs.

form is checked. This redirects to the appropriate branch of the guidelines. Cross-references among specific branches exist. For instance, some nominal, adjectival, and adverbial expressions are deverbal MWEs, e.g. (ro) ***aducere aminte*** (lit. 'bringing memory') 'reminding' is an AV stemming from the VID ***aduce aminte*** (lit. 'to bring memory') 'to remind'. The adjectival-adverbial diagram redirects to the verbal tests for such cases.

All in all, the language-independent part of the guidelines now contains 10 decision diagrams (7 of them for VMWEs) and 53 tests, 22 of which have equivalents in other diagrams. For instance the MORPH test from Fig. 1 has an equivalent in the VID, NMWE and AMWE branch. The remaining 8 branches contain 23 language-specific tests, most of them for MVCs. This shows a contribution of PARSEME to universalist modeling of language, understood as designing cross-linguistically consistent and applicable language descriptions.[7]

5. Challenges

Although the guidelines were carefully designed, we did face challenges while applying them. A recurrent problem for many languages concerns candidates like (i) *She did it in the absence of her son*, (ii) *She did it in his absence*. In Slavic languages, which exhibit noun inflection for case, the equivalent examples are (iii) (pl) *Zrobiła to pod nieobecność swojego syna* (lit. 'Did it under absence her son.GEN') 'She did it in the absence of her son', and (iv) (pl) *Zrobiła to pod jego nieobecność* (lit. 'She did it under his absence') 'She did in his absence'. The noun in (iii) takes no preposition, but subcategorizes for genitive case of its nominal complement *syn* 'son'. When annotating such examples, the first possibility is to consider that ***in the absence of*** in (i), and ***pod nieobecność*** (lit. 'under absence') 'in the absence' in (ii), have prepositional distribution (they commute with single prepositions *without* or *bez* 'without') and qualify as AdpIDs. On the other hand, ***in*** *…**absence*** in (ii), and ***pod*** *…**nieobecność*** (lit. 'under …absence') 'in …absence' in (iv), are adverbials and qualify as AdvIDs. However, this would ignore the fact that these examples actually contain the same expression. Another possibility is to say that only *in* and *absence* are lexicalized (i) and (ii), and the whole expression is an AdvID, which sucategorizes for the preposition *of* in (i). This might be satisfactory for English, but in Polish it would be artificial to label (iii) as an AdvID. Alternatively, we might choose solution 2 for languages that behave like English and solution 1 for those like Polish, but this would hinder the analysis of cross-lingual similarities. Currently, the two strategies co-occur across the corpus and coming up with a unified strategy is left for future work.

Cranberry items are used as the entrance test in several decision diagrams of the universal guidelines, like in the FuncMWE.1 test in Fig. 2. Those words are difficult to identify in non-spoken languages, like Ancient Greek. This is because firstly in order to rule out that an item truly is a cranberry item, we would need negative evidence, i.e., checking that the item does not appear in other contexts and constellations. In a non-spoken language, we can often not reconstruct what is lost. Furthermore, what is preserved may reflect certain clusters, e.g., texts that were of interest to a group of people wealthy enough to recopy texts and distribute them at a certain point in time or texts that happened to be discarded in a dry and hot environment rather than a damp and humid one (Hoffmann, 2005, ch. 8). Secondly, when we rely on smaller sub-samples, such as literary corpora in the *Thesaurus Linguae Graecae*, we risk overgenerating cranberry items (Fendel, 2025a). We have

[7] Instead of language universals (Greenberg, 1996).

solved the issue by allowing **NO or UNSURE** as an answer for tests involving cranberry items, as seen in Fig. 2, so that we can work our way down the decision diagram.

Annotation across sentence boundaries is currently not supported. This hinders appropriate representation of null anaphora, i.e. leaving out a direct object, e.g. in **λόγον ποιέομαι** (lit. 'to make word') 'to speak'. In Ancient Greek (Luraghi, 2003), null anaphora is a regular phenomenon, conversely to ellipsis, which is stylistically driven. Solving this issue is non-trivial, as it calls for re-engineering the CUPT file format (Sec. 7).

Testing for syntactic irregularity (via dedicated tests for VIDs and NMWEs) was complicated for Ancient Greek, partially by difficulty with ascertaining negative evidence (see above) and partially by the fact that the word order is primarily information-structurally driven. Thus, a fully fixed word order is highly indicative, e.g. in VID structures, but an information-structurally conceivable word-order variation is not indicative (Fendel, 2025b, ch. 3). We therefore had to rely on testing for other syntactic aspects, e.g. for valency, where possible.

6. Source texts and genres

The selection of texts to annotate followed various scenarios. For a majority of veteran languages, texts from edition 1.3 were re-annotated for non-verbal MWEs. In Polish, Romanian, Serbian and Slovene, all such texts were re-annotated, and in French, Greek, Hebrew and Swedish, only some of them. Persian and Portuguese are exceptions, since 2.0 texts are annotated from scratch.

A strong constraint for the veteran languages, for the first time in edition 2.0, was to also annotate a substantial quantity of new texts. This is because the PARSEME corpus is meant, inter alia, for evaluation campaigns of MWE-related tasks in NLP. Recently, such campaigns have been facing the problem of model contamination. Namely, the ubiquitously used large language models are trained on unrestricted web corpora, which likely contain any publicly available annotated dataset. No part of the previous PARSEME editions can thus be used for reliable evaluation in NLP.

New texts in veteran languages stem from: (i) Universal Dependencies (UD) treebanks (Zeman et al., 2025) for Greek, Persian, Polish, Portuguese and Swedish, (ii) *ELEXIS-WSD* corpus (Čibej et al., 2025) for Slovene and Serbian, (iii) Leipzig Corpora Collection (Goldhahn et al., 2012) for Hebrew, (iv) SELEXINI corpus (Scholivet et al., 2025) for French, (v) raw newspaper texts for Romanian and Serbian.

There are eight new languages in edition 2.0. The **Dutch** corpus draws upon two UD treebanks: Alpino and Lassy Small, containing newspaper texts from the Netherlands and from Belgium, covering both Dutch variants. The **Georgian** corpus could not rely on the Georgian UD treebanks, which are too small, but builds upon raw texts from Wikipedia, covering culture, education, geography, Georgia as a country, history, literature, philosophy, society and sports. This results in a ensures broad coverage and a focus on domains related to national identity. The **Latvian** corpus contains news, fiction, legal, spoken and academic texts from the UD LVTB treebank. The latter stems from automatic conversion of a pre-existing non-UD corpus with manual multilayered annotations – morphological, syntactic, and semantic (Gruzitis et al., 2018). The **Japanese** texts contain news and blog sentences randomly chosen from the UD GSD-LUW treebank. The **Ukrainian** corpus is based on the General Regionally Annotated Corpus of Ukrainian (GRAC),[8] with newspaper texts sourced from Ukrainian Week.

Finally, the last two newcomers, Egyptian and Ancient Greek, are non-spoken languages (or *corpus languages*). This category is represented in the PARSEME collection for the first time.

Egyptian is one of the earliest known written languages, with a written record spanning over 4,000 years and divided into 6 periods: Old Egyptian (ca. 2700–2000 BC), Middle Egyptian (ca. 2000–1550 BC), Classical Egyptian (from 1500 BC onwards), Late Egyptian (ca. 1550–700 BC), Demotic (7th Century BC to 5th century AD), and Coptic (4th century to 14th century AD). PARSEME currently covers Old Egyptian, with texts from the Egyptian Pre-Coptic treebank (Díaz Hernández and Carlo Passarotti, 2024). They come from Unas's and Teti's Pyramid Texts, initially recited during the mortuary rituals performed for these kings. The Pyramid Texts are a collection of spells recorded on the walls of the pyramids of Old Kingdom kings and queens.

The **Ancient Greek** corpus consists of texts in the Attic dialect dating from the 5th and 4th centuries BCE. These are courtroom speeches by the speechwriter Lysias (Speeches 1, 3, 7, and 12). As Athens was a direct democracy at the time, every individual needed to defend themselves in court, possibly with a helping hand from someone like Lysias, and speeches needed to be generally comprehensible (Willi, 2003). Courtroom speeches constitute therefore a comparatively well-resourced corpus variety.[9] The text selection complies with the one for Modern Greek (which includes parliamentary speeches) so as to

[8] https://uacorpus.org/en

[9] In a corpus language, the language variety qualifies as a corpus variety, i.e. the native speakers of a language are its texts (Fleischman, 2000).

facilitate comparative work.

egy*	el'	fa	fr*	grc*	he	ja'	ka	lv*	nl	pl	pt	ro	sl	sr*	sv	uk
95	89	99	98	20	36	76	82	74	94	62	42	78	63	71	47	54

Table 2: Inter-annotator agreement in terms of MWE-based F-score; (') self-agreement, (*) average among more than 2 annotators.

7. Quality and consistency

In order to ensure optimal quality, and consistency in particular, of the PARSEME corpus, we have developed a number of tools and follow a number of best practices. First, we strive to converge with another initiative sharing the same objectives of universality and diversity: the Universal Dependencies (de Marneffe et al., 2021), its treebank collection (Nivre et al., 2020) and the UDPipe parser (Straka, 2018). Because syntactic and semantic properties of MWEs are deeply intertwined, our corpus is released with underlying morphosyntactic annotations in a tabular CUPT format[10], which is an extension of UD's CoNLL-U[11] format. For each word, the 10 first columns contain UD-compatible segmentation and morphosyntactic annotation, and the 11th column bears MWE annotations from PARSEME. Updated UD corpus is released every 6 months. To optimize PARSEME/UD consistency, we have developed a synchronisation script which, for each sentence, searches its occurrences in the corresponding UD treebank, or parses it with UDPipe, and recopies the annotations from CoNLL-U to CUPT files. Tokenization conflicts, if any, are solved by a human expert in interactive mode.

Manual annotation also follows quality principles, even if we lack human resources to double-annotate and adjudicate all texts. An annotator goes through the sentences of the corpus one by one, identifies MWE candidates, and tests them against the guidelines. When automatically pre-annotated MWE candidates (from MWE lexicons, e.g. in Romanian or Georgian) are present, those are validated and other missed candidates are sought for. This annotation of running text is in contrast to some approaches cited in Sec. 2, where datasets build upon sentences pre-selected on the basis of their chances to contain particular MWEs.

Such a sequential annotation is still error prone. Therefore we use a custom tool for consistency checks, which, for every MWE annotated at least once, groups together all its positive and negative examples across the whole corpus. In this way, LLs can check that the same expression is consistently annotated and that literal or coincidental occurrences are not labeled. Consistency checks can greatly enhance the corpus quality. For instance, in Egyptian, Romanian and Polish, the F-measure (see below) between the running-text annotations and those corrected for consistency was 42, 86, and 92%, respectively.

We also enhanced the management of Gitlab repositories for individual languages with a continuous integration / continuous deployment (CI/CD) pipeline. Upon each push to the server, it checks the validity of all CUPT files present at the root of the repository. First, the latest UD validator is applied to the first 10 columns. Then, the MWE-related data are checked. The validation is done at 3 detailedness levels, from the backbone structure of a file, to the contents of each annotation and comment. The main branch of each repository is protected and used only for release. The LLs work in development branches and correct their data upon failed CI/CD pipelines, which significantly simplifies the release process.

Inter-annotator agreement (IAA) estimates how far two annotators can agree on the annotations of the same text, when working independently. Classical IAA measures like Cohen's κ include correction for chance agreement. In our case, where only some words are subject to annotation and their proportion in text is rather low, chance agreement is hard to estimate.[12] Instead, we use simply the MWE-based F-measure (Savary et al., 2017) between two annotators as IAA (see also Sec. 10). To estimate IAA, in most languages a fraction of the corpus containing from 100 (fa) to 5200 (he) sentences was annotated in parallel by two annotators. In French, Ancient Greek, Latvian and Serbian, more annotators were involved, and IAA is an average of the pairwise F-measures. In Egyptian, Greek and Japanese, where a single annotator was available, we measured self-agreement on the same text annotated after a period of time.

Tab. 2 shows that IAA is very disparate. It is quite high (above 70) for 9 languages. It should be stressed that the IAA files for the veteran languages were annotated from scratch rather than on top of pre-existing VMWE annotations. The case of French is special since the corpus builds upon pre-existing MWE annotations of all types and IAA measures only agreement of MWE categorization. IAA is the lowest in Ancient Greek, likely partly due to its non-spoken status (Sec. 5). The same effect is not seen in Egyptian, where IAA is measured as self-agreement. The two other cases of self-agreement (el, ja) are, expectedly, also high. The high score for Romanian is partic-

[10] https://gitlab.com/parseme/corpora/wikis/CUPT-format

[11] https://universaldependencies.org/format.html

[12] When only VMWEs were covered, chance agreement was estimated from the number of verbs.

	Sents	Tokens	MWEs	VMWE (%)	NMWE (%)	AMWE (%)	FuncMWE (%)	Deverb (%)	Idiom (%)
egy	2,182	22,540	633	220 (35)	145 (23)	87 (14)	181 (29)	56 (9)	257 (41)
el	2,732	61,580	1,245	627 (50)	201 (16)	185 (15)	232 (19)	13 (1)	482 (39)
fa	2,000	44,601	5,420	2,411 (44)	1,161 (21)	734 (14)	1,114 (21)	377 (7)	1,398 (26)
fr	4,084	98,525	5,636	1,343 (24)	1,979 (35)	1,314 (23)	1,000 (18)	59 (1)	2,569 (46)
grc	412	13,213	333	225 (68)	42 (13)	43 (13)	23 (7)	0 (0)	90 (27)
he	20,691	386,761	14,138	2,934 (21)	7,630 (54)	641 (5)	2,933 (21)	69 (0)	9,552 (68)
ja	1,963	36,817	3,447	1,781 (52)	612 (18)	317 (9)	737 (21)	39 (1)	879 (26)
ka	67,000	1,379,643	2,188	1,594 (73)	415 (19)	132 (6)	47 (2)	5 (0)	1,755 (80)
lv	13,784	225,731	3,203	751 (23)	555 (17)	925 (29)	972 (30)	74 (2)	737 (23)
nl	500	7,972	537	331 (62)	52 (10)	111 (21)	43 (8)	34 (6)	135 (25)
pl	26,266	429,231	13,704	7,990 (58)	970 (7)	3,107 (23)	1,637 (12)	50 (0)	2,091 (15)
pt	1,686	35,799	680	251 (37)	155 (23)	116 (17)	158 (23)	3 (0)	225 (33)
ro	71,220	1,423,424	64,601	18,084 (28)	8,752 (14)	14,672 (23)	23,093 (36)	1,123 (2)	14,766 (23)
sl	11,024	217,015	6,575	3,347 (51)	174 (3)	1,100 (17)	1,954 (30)	160 (2)	577 (9)
sr	10,468	212,650	9,555	3,635 (38)	3,509 (37)	946 (10)	1,465 (15)	171 (2)	3,915 (41)
sv	5,553	90,392	3,682	1,779 (48)	655 (18)	936 (25)	312 (8)	316 (9)	804 (22)
uk	12,078	198,555	5,993	2,804 (47)	818 (14)	1,017 (17)	1,354 (23)	345 (6)	1,134 (19)
All	253,643	4,884,449	141,570	50,107 (35)	27,825 (20)	26,383 (19)	37,255 (26)	2,894 (2)	41,366 (29)

Table 3: Overall corpus statistics per language: number of sentences, tokens, annotated MWEs, number and proportion (%) of verbal, nominal, adjectival/adverbial, functional, deverbal and idiomatic MWEs.

	egy	el	fa	fr	grc	he	ja	ka	lv	nl	pl	pt	ro	sl	sr	sv	uk	all
Tok/Sent	10.3	22.6	22.3	24.1	32.1	18.7	18.8	20.6	16.4	15.9	16.3	21.2	20.0	19.7	20.3	16.3	16.4	19.3
Tok/MWEs	35.6	49.1	8.2	17.5	39.7	27.4	10.7	630.5	70.5	14.8	31.3	52.6	22.0	33.0	22.3	24.5	33.1	35.2
Sent/MWEs	3.4	2.2	0.4	0.7	1.2	1.5	0.6	30.6	4.3	0.9	1.9	2.5	1.1	1.7	1.1	1.5	2.0	1.8

Table 4: Average sentence length (Tok/sent), token- (Tok/MWE) and sentence-MWE (Sent/MWE) ratios.

ularly encouraging, given that it accounts for over 50% of the total number of annotations across all languages (Tab. 3). Recall also that weak IAA can be mitigated at the stage of consistency checks. Some more investigations about the relatively low IAA in Swedish are offered by Stymne et al. 2026.

8. Statistics and split

The corpus statistics are available in Table 3, with proportions in parentheses. They cover the corpus size in number of sentences, tokens, and annotated MWEs. We also provide the distribution of MWEs according to their syntactic functions (VMWE, NMWE, AMWE and FuncMWE). While these 4 categories provide a partition over all annotated MWEs, two orthogonal categories describe the number and proportion of deverbal (NV and AV) and idiomatic (VID, NID, and AdjID)[13]MWEs.

There are substantial differences among the 17 covered languages. With only 412 sentences, Ancient Greek (grc) is the smallest corpus in PARSEME 2.0. With over 71,220 annotated sentences, the Romanian corpus is the largest. In terms of annotations, Polish, Romanian and Hebrew have more than 10,000 MWEs, whereas Ancient Greek, Egyptian, Dutch (nl) and Portuguese (pt) have less than 1,000.

Looking at the four main MWE categories, we see that, on average, VMWEs dominate (35% of all MWEs), FuncMWEs follow (26%), and NMWEs and AMWEs have similar frequencies (20% and 19%). However, there is variability across languages: with only 2% of FuncMWEs and 73% of VMWEs, Georgian (ka) is an outlier. Other languages such as Egyptian (egy) or Portuguese (pt) have proportions very close to the global average.

Table 4 indicates that sentence length is quite homogeneous, with around 20 tokens per sentence, except for Egyptian with 10.3, and Ancient Greek with 32.1. Overall, the number of sentences correlates with the number of MWEs, but Georgian (ka) contains only 2,188 MWEs for 67,000 annotated sentences, or an average of one MWE every 30.6 sentences. In comparison, Romanian contains 64,601 MWEs for 71,220 sentences, or an average of one MWE every 1.1 sentence. Five languages, Greek (el), Japanese, Dutch (nl), Persian (fa) and Ukrainian (uk), contain more than 1 MWE

[13]These 3 categories seem to exhibit overall the highest semantic non-compositionality.

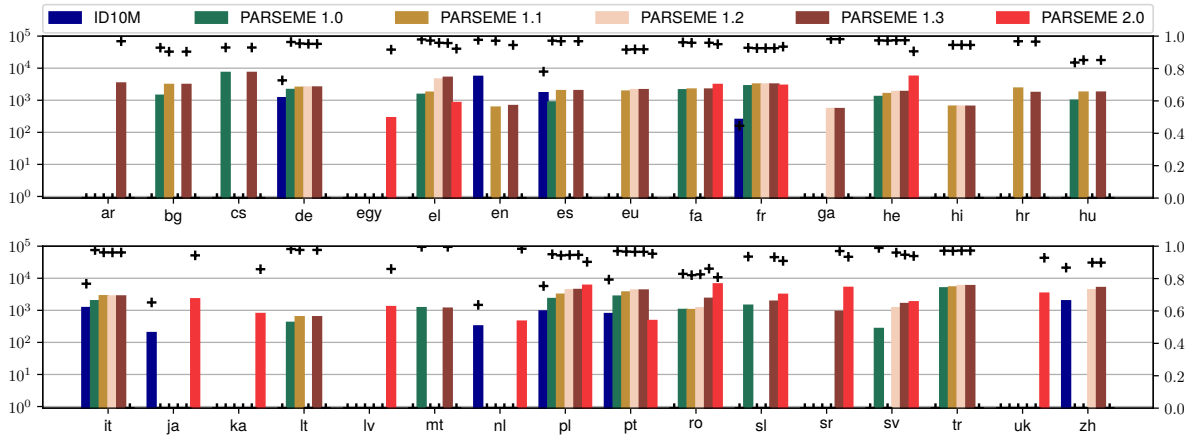

Figure 3: Diversity of MWE forms per language in ID10M and different PARSEME versions. Richness on the left axis (higher is better), represented by the large colored bars. Scale is logarithmic. Shannon Evenness on the right axis (higher is better), represented by the black crosses.

per sentence on average. In total, there is an average of 1 MWE for every 1.8 sentences. This variability could stem from limited cross-lingual suitability of the guidelines, their different interpretations, text genres with different MWE densities, or language-specific features.

The PARSEME corpus is notably meant to develop and benchmark MWE identification systems. Therefore, it is split into TRAINing, DEVelopment, and TEST parts. We distinguish: (i) OLD texts with previously published MWE annotations (e.g., from previous PARSEME releases) (ii) NEW texts. Due to contamination issues (Sec. 6), TEST sets rely on NEW texts only. We assume that for evaluation to be reliable, TESTS should contain at least 500 MWEs per language. Some TESTS contain slightly more than 500 MWEs if the last selected sentence has more than one MWE. For two small languages, Ancient Greek (grc) and Dutch (nl), the TESTS contain 333 and 434 MWEs, respectively.

To create a TEST for a language, we first set aside all NEW sentences containing exactly one MWE of type VID, NID, or AdjID.[14] If there are more than 150 such sentences, only 150 (randomly selected) are added to TEST. Otherwise, they are all added. The TEST is then completed with randomly selected sentences from the remaining part of NEW texts. All sentences not selected for TEST are split with 90%/10% ratio between TRAIN and DEV. Ancient Greek, with fewer than 500 annotated MWEs, has no TRAIN/DEV datasets.

9. Diversity

To estimate in-text diversity (Estève et al., 2025), we borrow from the fields of ecology (Patil and Taillie, 1982), economy (Stirling, 2007), and information networks (Ramaciotti Morales et al., 2021), where diversity is relatively well formalized. We define *categories* (whose diversity is to be quantified) as distinct MWE forms annotated in the corpus. *Elements* are all the occurrences of these categories. We select two diversity measures. *Richness* is simply the number of categories. *Shannon evenness* (Ramaciotti Morales et al., 2021) measures balance of the categories' distribution. Low balance usually means that few forms are very frequent and many forms are rare. All other things being equal, the higher the richness the higher the diversity. The same holds for balance.

Fig. 3 shows the diversity in the different editions of the PARSEME corpus – also including languages which are not released in edition 2.0 – as well as in the ID10M dataset (Tedeschi et al., 2022) for comparison (Sec. 2). Richness is represented by the large colored bars and the left-side axis, whose scale is exponential. We observe that richness grew in PARSEME 2.0 wrt. the previous editions in Persian, Hebrew, Polish, Romanian, Slovene, Serbian and Swedish. Balance is rendered by crosses and the (linear) right-hand axis. For all PARSEME editions balance exceeds 0.8 in all languages.

	de	en	es	fr	it	ja	nl	pl	pt	zh
ID10M	726	1,202	651	939	818	211	548	862	764	244
1.0	144	-	159	486	427	-	-	220	414	-
1.1	173	124	182	528	430	-	-	274	638	-
1.2	173	-	-	525	430	-	-	396	728	649
1.3	173	124	182	525	430	-	-	396	728	820
2.0	-	-	-	98	-	36	7	429	35	-

Table 5: Number of tokens (in thousands) per language per dataset in ID10M and PARSEME.

[14] Meant for a future corpus of MWE paraphrases.

		egy	fa	fr	grc	he	ja	ka	lv	nl	pl	pt	ro	sl	sr	sv	uk
MWE-based	P	0	54.9	41.4	18.0	24.7	22.4	5.4	17.3	80.0	25.7	11.1	33.1	26.6	32.3	27.9	28.8
	R	0	40.7	32.6	40.0	38.5	16.9	63.1	54.1	50.0	23.4	30.7	33.2	41.9	30.4	41.5	43.3
	F	0	46.8	36.5	24.9	30.1	19.3	10.0	26.3	61.5	24.5	16.3	33.2	32.6	31.3	33.3	34.6
token-based	P	5.4	62.6	54.4	23.8	28.5	31.9	5.2	18.7	75.0	30.8	20.2	40.0	33.2	40.5	33.4	35.1
	R	6.0	61.8	52.8	63.6	46.0	39.1	74.2	66.2	56.2	32.9	67.8	51.9	60.9	47.0	69.8	57.9
	F	5.7	62.2	53.6	34.7	35.2	35.1	9.8	29.1	64.2	31.8	31.1	45.2	43.0	43.5	45.2	43.7

Table 6: Baseline scores in MWE identification using gpt-5-mini.

ID10M has higher richness and balance for English, but lower for the other 9 languages it covers,[15] which shows PARSEME's contribution to less-resourced languages. This is particularly noteworthy wrt. Tab. 5. Namely, richness usually strongly correlates with corpus size. But richness in PARSEME is mostly substantially higher than in ID10M, despite its size being substantially lower, except in Chinese. Two reasons might be that ID10M covers MWEs from Wiktionary only and its annotation is semi-automatic, while in PARSEME running text is annotated manually.

10. Baselines in MWE identification

We performed first experiments with our corpus to test generic LLMs in the task of MWE identification (Constant et al., 2017). We implemented a system available open source[16] based on API calls for interaction with LLMs. It allows communication with public cloud-based APIs (such as xAI or OpenAI) or local models hosted in Ollama. We used succinct MWE and task definitions in user and system prompts (see App. 15.1). The system was not in contact with the training data. Tab. 6 shows the results of the best performing *gpt-5-mini model*, in terms of MWE-based and scores (Savary et al., 2017), which account for exact and partial MWE matches, respectively. Expectedly, token-based scores are overall higher than MWE-based ones. Precision is considerably higher than recall for some languages (fa, fr, ja, nl) and vice-versa for some others (grc, he, ka, lv, pt, sl, sv, uk), with Georgian (ka) showing the most striking difference. Reasonable overall performances are achieved for Dutch with a particularly high precision. The token-based F-score for Persian and French are comparable to Dutch but their MWE-based scores are much lower. For all other languages, the F-scores are low to very low, to the point of not exceeding 10.0 in Georgian and Egyptian, and even yielding F_{MWE}=0 in Egyptian, where no MWE was correctly identified. We also tested other LLMs (LLaMA3.3 and Qwen3), but their results were considerably lower. These results confirm (cf. Sec. 1) that generic LLMs do not encode sufficient knowledge about MWEs and have to be fine-tuned to solve MWE-related tasks, which shows the importance of MWE-annotated corpora.

[15] For Portuguese, richness in ID10M is higher than in PARSEME 2.0 but lower than in previous editions.

[16] https://github.com/racai-ai/mwe_baseline/

11. Conclusions and future work

Edition 2.0 of the PARSEME corpus is an important milestone in the long-standing efforts of the PARSEME community towards universality-driven modeling of idiomaticity. For the first time we cover all categories of MWEs, not only verbal ones. Having annotated texts in 17 languages from different language genera, with relatively little intervention of language-specific tests, corroborates the strong universality of the guidelines. The VMWE categories previously considered universal: LVCs and VIDs, are found in all 7 new languages. Newly covered categories: NMWEs, AMWEs and FuncMWEs were also found in all 17 languages. We hope to gain finer linguistic insights from future corpus studies. Also for the first time we cover non-spoken languages: Egyptian and Ancient Greek. They provide a extra validation of the overall framework but bring new challenges.

PARSEME is an open-ended project. We hope that the 16 languages released in edition 1.3, but not in 2.0, will extend their annotations to non-verbal MWEs, and that new languages will join. We plan to further enhance the annotation guidelines, notably of the MVC branch, where relatively little cross-lingual commonalities have been captured. The low IAA score for some languages calls for more attention. We might redo the IAA estimations now, in the post-annotation phase, and discuss recurrent issues arising in the training of annotators and in the understanding of the guidelines. Other perspectives include automating the tests from the guidelines to support manual annotation and automatic MWE identification. Finally, closer convergence with Universal Dependencies is envisaged, for the sake of better understanding of idiomaticity and of its interactions with syntax.

12. Ethics statement

In this work we strive towards a better inclusiveness of world's languages in language resources and tools, via building a community of experts working towards a common goal of a unified understanding and modeling of idiomaticity. Previous editions of the PARSEME corpus covered 26 languages and 7 new ones join in this edition. Despite this increasing coverage, our dataset does not escape the global tendency in NLP of being biased towards Indo-European languages. Among the 17 languages of edition 2.0, only 4 are not Indo-European: Egyptian, Georgian, Hebrew and Japanese.

We can also estimate how well we account for lower-resourced languages using the taxonomy proposed by Joshi et al. (2020). As seen in Tab. 7, even if we do not focus on the *Winners* in the first place, we do dedicate most efforts to the *Underdogs* and the *Raising Stars*. Only one our (upcoming) language is among the *Hopefuls* and we have no contributions to those languages which suffer most from the non-diversity biases in NLP: the *Scraping-Bys* and the *Left-Behinds*. One important contribution though is to cover non-spoken languages: Egyptian and Ancient Greek, which are not considered in this typology.

Joshi class	Class name	PARSEME 2.0 languages
	Ignored by Joshi et al. (2020)	egy, grc
0	Left-Behinds	
1	Scraping-Bys	
2	Hopefuls	mr (upcoming)
3	Rising Stars	el, he, ka, lv, ro, sl, uk
4	Underdogs	fa, nl, pl, pt, sr, sv
5	Winners	fr, ja

Table 7: Languages in the PARSEME 2.0 corpus according to the resourcedness status by Joshi et al. (2020)

Each universalist approach, like ours, tends to unify different points of view on the same phenomena. While this leads us towards descriptions that allow cross-language comparability, it may also impoverish valid intuitions and models which exist before the unification effort. This itself may have an inverse effect on diversity than the intended one. Even if this risk is mitigated by inclusiveness principles described in Sec. 3, it remains very challenging for many language communities to make their languages' phenomena represented on a par with better resourced languages.

13. Limitations

Additionally to the limited inclusiveness issues discussed in Sec. 12, this work has important limitations related to uneven inter-annotator agreement across languages (Sec. 7), which might stem either from insufficient preparedness of some annotators, or imperfections of the annotation guidelines and methodology. Continuous enhancements of the corpus are therefore necessary, but difficult for such a big community.

Linguistic work showed that semantic non-compositionality is a matter of scale rather than a binary phenomenon. In PARSEME, for the reasons of tractability, we simplify this problem by drawing a line between non-compositional expressions on the one hand and compositional ones the other. This is made possible by the determinism of the annotators' decisions (up to the answers to the atomic tests), so that any MWE candidate, notably from a gray zone of non-compositionality, is always assigned to the same class. However this binary model might not be straightforwardly applicable to linguistic studies. Therefore, our typology should certainly not shadow more fine-grained and accurate classifications, especially when language-specific phenomena are concerned.

Another limitation lies in the unresolved discrepancies occurring in the unified annotation guidelines (Sec. 5). Future editions of the annotation guidelines should address these issues.

Finally, looking at the baseline results in MWE identification (Sec. 10), it should be kept in mind that they stem from preliminary studies, using unsophisticated LLM prompts. Therefore, care should be taken when discussing the knowledge about MWEs that LLMs might or might not encode.

14. Acknowledgements

We are grateful to all the annotators having participated in the project (see App. 15.2).

This work received support from the CA21167 COST action UniDive, funded by the European Union via the COST (European Cooperation in Science and Technology). Further support came from: (1) French Agence Nationale pour la Recherche, via the SELEXINI project (ANR-21-CE23-0033-01), (2) Swedish national research infrastructure Språkbanken, jointly financially supported by the Swedish Research Council (2025–2028; grant 2023-00161) and the 10 participating partner institutions, (3) Brazilian National Council for Scientific and Technological Development (CNPq 404722/2024- 5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG), (4) Latvian Council of Science via the project "Advancing Latvian com-

putational lexical resources for natural language understanding and generation"(LZP2022/1-0443), (5) Slovenian Research and Innovation Agency (research core funding No. P6-0411 *Language Resources and Technologies for Slovene* and No. P6-0215 *Slovene Language – Basic, Contrastive, and Applied Studies*), (6) the Ministry of Science, Technological Development and Innovation, Republic of Serbia (GRANT 451-03-33/2026-03/200174) and the Science Fund of the Republic of Serbia #7276, *Text Embeddings-Serbian Language Applications, TESLA*; (7) the Stichting Taaltechnologie Utrecht University, (8) the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them, (9) a grant of the Ministry of Research, Innovation and Digititalization - UEFISCDI, Romania, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (10) NATO Science for Peace and Security Programme under grant id. G8648, project DeepNewsDef, (11) a grant of the Ministry of Education and Research, CCCDI - UEFISCDI, Romania, project number PN-IV-P8-8.2-NATO-SPS-2025-0005, within PNCDI IV.

15. Appendices

15.1. Appendix A. Prompting Configuration

To ensure experimental reproducibility, we document below the system and user prompts used in LLMs experiments, together with decoding parameters and execution setup.

15.1.1. System Prompt

The following system prompt was used consistently across all experimental splits:

System Prompt

"You are a linguist that identifies multiword expressions. Multiword expressions are idiosyncratic interpretations that cross word boundaries. Named entities are not multiword expressions. Do not reveal chain-of-thought, internal reasoning, or ‹think› sections. Only return the final answer."

This prompt establishes the linguistic role of the model as an expert annotator and provides an operational definition of MWEs. It explicitly excludes named entities from consideration and constrains the model to produce only the final output without revealing intermediate reasoning or chain-of-thought content.

15.1.2. User Prompt

The following user prompt template was used for each input sentence:

User Prompt Template:

"Multiword expressions are idiosyncratic interpretations that cross word boundaries. Named entities are not multiword expressions.

Task: You are an expert annotator for the PARSEME task. Extract multiword expressions (MWEs) from the sentence below.

Definition (apply all): An MWE is a fixed or semi-fixed sequence of ≥2 tokens whose meaning or usage is not freely compositional (idioms), or is a conventionalized unit (light-verb constructions, verb-particle constructions, fixed/prepositional phrases, lexicalized compounds, etc.).

Output format: If none found, output exactly: NONE. Otherwise, output only the MWEs separated by a single pipe character: MWE1|MWE2|... Keep the original language and do not translate into English. %s "

The user prompt was identical across the experimental splits and no split-specific instructions or gold labels were included. This ensured that the model did not receive information that could inadvertently bias the evaluation.

15.2. Appendix B. Annotators' teams

EGY: Roberto Antonio Díaz Hernández (LL),[17] Carmen Romero Fernández; **EL**: Voula Giouli (LL), Georgia Karamousli, Stella Markantonatou, Athanasios Pantazis, Eri Papanikolaou, Styliani Tononidou; **FA**: Mehrnoush Shamsfard (LL), Vahide Tajalli (LL), Kosar Seraji; **FR**: Takuya Nakamura (LL), Mathilde Deletombe, Manon Scholivet; **GRC**: Victoria Fendel (LL), Sirui Cheng, Artemis Ffytche, Alexandre Guilloteau, Svetlana Ó Siochfhradha Prešern, Paraskevi Platanou, Elena Squeri, Eirene Vlastou; **HE**: Chaya Liebeskind (LL), Yael Dayan, Elinoy Damari; **JA**: Takuya Nakamura (LL); **KA**: Irina Lobzhanidze (LL), Svetlana Berikashvili; **LV**: Gunta Nešpore-Bērzkalne (LL), Agute Klints, Laura Rituma; **MR**: Aakanksha Padhye (LL); **NL**: Carole Tiberius (LL), Lut Colman; **PL**: Agata Savary (LL), Anna Latusek, Alina Wróblewska; **PT**: Adriana Pagano (LL), Gabriela Berndt de Souza, Leticia Guedes, Eric Tempesta; **RO**: Verginica Barbu Mititelu (LL), Ioana Biolan, Ioana Buhnila, Elena Irimia, Mihaela Cristescu, Cătălin Mihăilă, Isabella Șinca, Amalia Todirașcu, Carmen Vasile; **SL**: Jaka Čibej (LL), Polona Gantar (LL), Simon Krek, Mija Bon, Kaja Dobro-

[17] LL stands for language leader.

voljc, Nives Hüll; **SV**: Sara Stymne (LL), Astrid Berntsson Ingelstam, Stella Lundqvist, Ida Nilsson, Eva Pettersson; **SR**: Cvetana Krstev (LL), Ranka Stanković (LL), Aleksandra Marković (LL), Marina Bagi, Milica Dinić Marinković, Marijana Đukić; **UK**: Olha Kanishcheva (LL), Maria Shvedova, Hanna Sytar.

16. Bibliographical References

Kübra Adalı, Tutkum Dinç, Memduh Gokirmak, and Gülşen Eryiğit. 2016. Comprehensive annotation of multiword expressions for turkish. In *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*, pages 60–66, Konya, Turkey.

Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2024. Training corpus SUK 1.1. Slovenian language resource repository CLARIN.SI.

Kfir Bar, Mona Diab, and Abdelati Hawwari. 2014. Arabic multiword expressions. In *Language, Culture, Computation. Computational Linguistics and Linguistics: Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday*, pages 64–81. Springer.

Eduard Bejček and Pavel Straňák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1–2):7–21.

Gosse Bouma, Jan Odijk, and Carole Tiberius. 2024. Towards a Dutch Parseme corpus. In *Proceedings of the second general UniDive meeting*, pages 29–37, Pisa, Italy. Unidive Project.

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden. Association for Computational Linguistics.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2):415–479.

Jaka Čibej, Simon Krek, Carole Tiberius, Federico Martelli, Roberto Navigli, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Rob Tempelaars, Rute Costa, Ana Salgado, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, and Sia Kolkovska. 2025. Parallel sense-annotated corpus ELEXIS-WSD 1.3. Slovenian language resource repository CLARIN.SI.

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

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Roberto Antonio Díaz Hernández and Marco Carlo Passarotti. 2024. Developing the Egyptian-UJaen treebank. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 1–10, Hamburg,Germany. Association for Computational Linguistics.

Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 337–344, Trento, Italy. Association for Computational Linguistics.

Victoria Beatrix Fendel. 2025a. Zenodo. [link].

Victoria Beatrix Fendel. 2025b. *Giving gifts and doing favours: Literary classical Attic Greek support-verb constructions*. Brill, Leiden and Boston.

Victoria Beatrix Fendel and Matthew Ireland. 2023. Discourse cohesion in xenophon's on horsemanship through sketch engine. *Digital Humanities Quarterly*, 17(3):n/a.

Suzanne Fleischman. 2000. Methodologies and ideologies in historical linguistics: On working with older languages. In Susan Herring, Pieter Reenen, and Lene Schøsler, editors, *Textual parameters in older languages*, pages 33–58. John Benjamins, Amsterdam.

Voula Giouli, Foufi Vassiliki, and Angeliki/Aggeliki Fotopoulou. 2019. Annotating greek vmwes in running text: a piece of cake or looking for a needle in a haystack? In *Proceedings of the 13th International Conference on Greek Linguistics*, London, UK.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Joseph Harold Greenberg, editor. 1996. *Universals of language*. MIT Press.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 90:57–71. Paris : Larousse.

Maurice Gross. 1986. Lexicon-grammar: The representation of compound words. In *Proceedings of the 11th Conference on Computational Linguistics*, COLING '86, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. 2018. Creation of a balanced state-of-the-art multilayer corpus for nlu. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 4506–4513.

Sebastian Hoffmann. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-Based Study*. Taylor and Francis, Florence, US.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

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

Heiki-Jaan Kaalep and Kadri Muischnek. 2008. Multi-word verbs of estonian: a database and a corpus. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 23—26, Marrakech, Morocco.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008a. A French Corpus Annotated for Multiword Expressions with Adverbial Function. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51, Marrakech, Morocco.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008b. A French Corpus Annotated for Multiword Nouns. In *Proceedings of the Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions*, pages 27–30, Marrakech, Morocco.

Irina Lobzhanidze. 2019. Computational model of the modern georgian language and search patterns for an online dictionary of idioms. In *Language, Logic, and Computation*, pages 187–208, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on mwe resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Silvia Luraghi. 2003. Definite referential null objects in ancient greek. *Indogermanische Forschungen*, 108:167–194.

Stella Markantonatou, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary, and Veronika Vincze. 2021. PMWE conventions for examples containing multiword expressions. Technical report, Phraseology and Multiword Expressions – book series at Language Science Press.

Stella Markantonatou and Niki Samaridi. 2017. Revisiting the grammatical function "object" (obj and objθ). In Stella Markantonatou and Manfred Sailer, editors, *Multiword expressions: Insights from a multi-lingual perspective*, pages 187–213. Language Science Press, Berlin.

Filip Miletić and 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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Thiago Pardo, Magali Duran, Lucelene Lopes, Ariani Felippo, Norton Roman, and Maria Nunes. 2021. Porttinari - a large multi-genre treebank for brazilian portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 1–10, Porto Alegre, RS, Brasil. SBC.

Ganapati P. Patil and Charles Taillie. 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association*, 77(379):548–561. Number: 379 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Lauma Pretkalniņa. 2023. *Formāls latviešu valodas gramatikas modelis un tā realizācija mašīnlasāmā sintakses korpusā*. Ph.D. thesis, Latvijas Universitāte.

Lauma Pretkalniņa, Laura Rituma, and Baiba Saulite. 2018. Deriving enhanced universal dependencies from a hybrid dependency-constituency treebank. In *Text, Speech, and Dialogue*, volume 11107, pages 95–105. Springer.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115. Publisher: Elsevier.

Carlos Ramisch, 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, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, 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, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task

on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. 2016. Mwes in treebanks: From survey to guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023a. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

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

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023b. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Manon Scholivet, Agata Savary, Louis Estève, Marie Candito, and Carlos Ramisch. 2025. SELEXINI – a large and diverse automatically parsed corpus of French. In *Proceedings of the*

18th Workshop on Building and Using Comparable Corpora (BUCC), pages 83–98, Abu Dhabi, UAE. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Mojgan Seraji, Filip Ginter, and Joakim Nivre. 2016. Universal dependencies for persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2361–2365.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Sara Stymne, Astrid Berntsson Ingelstam, and Eva Pettersson. 2026. Swedish Multiword Expression Corpora in PARSEME. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Veronika Vincze and János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1110–1118, Beijing, China. Coling 2010 Organizing Committee.

Andreas Willi. 2003. *The languages of Aristophanes: aspects of linguistic variation in Classical Attic Greek*. Oxford University Press, Oxford.

17. Language Resource References

Carlos Ramisch, Silvio Ricardo Cordeiro, et al. 2018. Annotated corpora and tools of the PARSEME shared task on automatic identification of verbal multiword expressions (edition 1.1). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Carlos Ramisch, Bruno Guillaume, Agata Savary, et al. 2020. Annotated corpora and tools of the PARSEME shared task on semi-supervised identification of verbal multiword expressions (edition 1.2). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, et al. 2017. Annotated corpora and tools of the PARSEME shared task on automatic identification of verbal multiword expressions (edition 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Agata Savary, Carlos Ramisch, Bruno Guillaume, et al. 2023. PARSEME corpora annotated for verbal multiword expressions (version 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2025. Universal dependencies 2.16. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).