Corpora, Annotation systems, and NLP Tools

The Romanian Component of the Parallel ELEXIS Corpus: ELEXIS-Ro

Verginica Barbu Mititelu
RACAI, Bucharest
vergi@racai.ro

Elena Irimia
RACAI, Bucharest
elena@racai.ro

Simina Popa
University of Bucharest
popa.simina2005@gmail.com

Mihaela Cristescu
University of Bucharest
mihaela.cristescu@litere.unibuc.ro

Cătălin Mihăilă
RACAI, Bucharest
catalin@racai.ro

Lea Radu
The Saint Sava National College
leacristiana@icloud.com

Ioana Bucșa
Gheorghe Șincai National College
ioana.bucsa34@gmail.com

Abstract

We present the steps taken to extend the ELEXIS parallel corpus to include Romanian. This extension is based on automatic processing that is followed by manual validation and correction. The English original corpus was translated into Romanian, tokenized and then UDPipe was run on it. The resulting part-of-speech tagging was manually validated and some remarks on the types of mistakes encountered are presented herein. The corpus was further automatically enriched with semantic annotation of the types multiword expressions of all parts of speech and named entities. The multiword expressions were labeled with the types defined in the PARSEME project. Three types of named entities were annotated: person, organization and location. A double validation and correction process took place and its results are also presented here. The paper also includes statistics of the corpus: number of tokens per part of speech, number of multiword expressions per type and of name entities.

Keywords: parallel corpus, Romanian, translation, tokenization, PoS tagging, lemmatizaton, manual validation, multiword expressions, named entities.

1 Introduction

Parallel corpora – collections of texts aligned across two or more languages – are fundamental resources for multilingual research and language technology, enabling systematic analysis of how meaning, structure, and discourse are transferred across languages, while revealing translation patterns, equivalence relations, and cross-linguistic variation. When carefully produced and validated, they serve as reliable repositories of linguistic knowledge and remain indispensable even in the era of Large Language Models (LLMs), whose performance still depends on high-quality multilingual data; in this context, parallel corpora provide transparent, verifiable alignments for evaluation, fine-tuning, and interpretability. Their value is further enhanced by the inclusion of low-resourced languages, which supports linguistic diversity, reduces technological inequality, and enables cross-lingual transfer, as well as by multi-layer linguistic annotation (morphological, syntactic, semantic, and discourse), which allows deeper analysis of language correspondences and supports advanced applications such as machine translation, cross-lingual retrieval, terminology extraction, and knowledge discovery.

We present an extension of the ELEXIS parallel corpus (Martelli et al., 2021) through the inclusion of Romanian, thus enhancing its linguistic coverage and research potential. As a fourth Romance language in the dataset, Romanian enables richer comparative and contrastive studies within the Romance family, supporting analyses of lexical, syntactic, semantic, and discourse-level variation. Its addition is valuable also from a typological and areal perspective, given Romanian's distinct evolution in Eastern Europe and its extensive contact with Slavic and Balkan languages. Together with Bulgarian, also included in the corpus, this allows for the investigation of Balkan Sprachbund phenomena – such as definiteness marking, analytic constructions, and clitic doubling – while supporting research across genealogical, typological, and contact-linguistic dimensions.

The ELEXIS corpus is a manually annotated multilingual parallel dataset designed for lexicography and Word Sense Disambiguation (WSD), originally introduced in 2021 with ten European languages and consisting of 2,024 validated parallel sentences derived from WikiMatrix. It features five layers of annotation: tokenisation, subtokenisation, lemmatisation, part-of-speech tag-

ging, and WSD. It aims to advance semantic annotation and improve sense inventories. In its 2.0 version[1], other annotation types (such as named entities) are also available for some languages

Against the landscape of the extending the ELEXIS corpus with other languages (see Section 2) we paint the picture of the work done for adding Romanian (Section 3) and present statistics relevant to the annotation levels present in the corpus, before envisaging future work (Section 6), announcing corpus availability (Section 8) and concluding the paper (Section 7).

2 Related Work

The ELEXIS project[2] resulted in a multilingual parallel corpus with sense annotation across ten languages, later extended within the UniDive COST Action (Savary et al., 2024). One major component is the Danish corpus (DA-ELEXIS) (Pedersen et al., 2023), among the largest sense-annotated resources for Danish, featuring multiple annotation layers (tokenisation, sub-tokenisation, lemmatisation, and POS tagging) generated under Universal Dependencies (UD) guidelines[3] and manually validated. The corpus comprises over 32,000 tokens with full sense annotation for content words, supported by a semi-automated annotation tool. Using DanNet as the primary sense inventory, the project also addressed compounds through splitting strategies and encoded multiword expressions (MWEs) with unified semantic labels. Inter-annotator agreement reached 0.78 after guideline refinement, indicating reliable semantic annotation.

The Serbian extension (Krstev et al., 2024) focuses on semantic annotation and multilingual alignment, particularly for MWEs and named entities (NEs). Sentences from WikiMatrix were machine-translated into Serbian and manually validated, with attention to translation errors and structural inconsistencies. Annotation layers (tokenisation, lemmatisation, POS tagging) were automatically generated and manually checked. The study highlights cross-linguistic variation by showing how MWEs differ structurally across languages, including cases where phrases correspond to single words. The dataset includes hundreds of nominal and verbal MWEs, as well as named entities identified through both lexical resources and NER tools, enabling iterative expansion of the Serbian lexical repository.

For semantic annotation, the Serbian dataset relies primarily on Serbian WordNet (SrpWN), aligned with Princeton WordNet (Fellbaum, 1998) through definition-based matching due to the absence of an interlingual index. Automatic translation of synonym sets and integration of additional lexical resources supported this alignment process. Results show that only about 39% of MWEs were directly matched in SrpWN, reflecting significant cross-linguistic variation and underscoring the challenges of multilingual semantic alignment, particularly in the treatment of MWEs and lexical equivalence across languages.

3 Work Methodology

This section outlines the integration of Romanian into the ELEXIS corpus, including the translation of the English source texts, tokenization, and morpho-syntactic annotation. Following the established ELEXIS methodology, these processes are initially carried out automatically using specialized tools for machine translation and linguistic annotation, and are subsequently manually revised and validated to ensure accuracy, cross-linguistic consistency, and compatibility with the multilingual framework.

3.1 Language Processing

3.1.1 Translation

Automatic translation from English into Romanian was performed using Google Translate in February 2024, with subsequent manual validation showing that 1,437 out of 2,024 sentences (around 70%) required no correction, while the remaining 30% were manually revised. Most machine translation (MT) errors fall into recurring, predictable categories, the most frequent being lexical and register-related issues. These include contextually inappropriate or stylistically neutral synonyms, incorrect terminological choices, and literal or non-idiomatic expressions that mirror source-language structures. Human revision plays a key role in selecting contextually appropriate equivalents, ensuring correct register (technical, academic, or colloquial), and normalizing collocations to align with natural Romanian usage.

A second major category involves morpho-syntactic and structural errors, where MT outputs reflect source-language interference or incorrect

[1] https://www.clarin.si/repository/xmlui/handle/11356/2101.

[2] https://project.elex.is/

[3] https://universaldependencies.org/

grammatical relations. These include agreement errors, improper determiners, incomplete or ambiguous constructions, and unnatural word order, all of which require systematic correction to ensure grammaticality and fluency in Romanian. Additionally, issues related to named entities (NEs) and linguistic conventions are observed, such as incorrect capitalization, improper adaptation of proper names and titles, and inadequate localization of institutional names and acronyms.

Overall, while MT provides semantically adequate translations, it tends to preserve source-language structures, resulting in outputs that lack naturalness and stylistic appropriateness. Human intervention is therefore essential to refine lexical choices, adjust syntactic patterns, and ensure coherence with target-language norms, ultimately producing high-quality translations suitable for inclusion in a linguistically reliable parallel corpus.

3.1.2 Tokenisation

Automatic tokenisation of the Romanian ELEXIS component was performed using UDPipe (Straka et al., 2016) with a model trained on the Romanian Reference Treebank (RRT) (Barbu Mititelu, 2018), after which the output was manually corrected in a shared Google Sheets environment following unified multilingual guidelines. These guidelines stress accurate tokenisation as the basis for all envisioned subsequent annotation layers (lemmatisation, POS tagging, UD parsing, MWE and NE annotation, and WSD), and specify that ad-hoc compounds are the only multiword units systematically split for token-level disambiguation in Romanian.

Romanian-specific tokenisation rules were adapted from previous UD-based resources and detailed treatments for hyphenated structures were defined. Established dictionary compounds and toponyms are kept as single tokens, while semi-productive or context-specific compounds (e.g. ad-hoc formations) are split and annotated by span. Additional rules govern ambiguous hyphenation cases, including foreign compounds, inflectional hyphens, numerical intervals, and lexicalised expressions, ensuring consistent handling across diverse morphological and orthographic phenomena.

A major focus is the treatment of Romanian cliticisation and vowel elision, where hyphens and apostrophes signal phonological reduction in pronouns (e.g., *spune-mi* "tell me", where the initial vowel *î* from the pronoun *îmi* is elided), auxiliaries (e.g., *cărțile-s noi* "the books are new" instead of *cărțile îs*[4] *noi*), prepositions (e.g., *c-un cuțit* "with a knife", where the preposition *cu* is shortened to *c-*), and conjunctions (e.g., *c-a venit* "that he came", where the conjunction *că* "that" is reduced to *c-*); these are systematically split with the clitic attached to the appropriate host. Further rules cover symbols, formulas, units, currency, slashes, named entities, abbreviations, and IT terms, balancing fine-grained segmentation with the preservation of lexical integrity to ensure consistency across linguistic categories.

3.1.3 Lemmatisation and part-of-speech tagging

Part-of-speech tagging, morphological feature assignment, and dependency parsing for the Romanian ELEXIS corpus were initially performed automatically using UDPipe (Straka, 2018), trained on the Romanian Reference Treebank (RRT), and exported in CONLL-U format before conversion to TSV3 for compatibility with INCEpTION (Klie et al., 2018). The resulting annotations were subsequently manually reviewed and corrected at the lemmatization and morphological levels, ensuring consistency and accuracy across the dataset.

During manual validation, several recurrent error types were identified, largely driven by lexical and morphological homonymy. A major source of inconsistency involved ambiguity between adjectives and participles, as well as between proper nouns and common nouns in capitalized or context-dependent uses (e.g. *Pământ* "Earth" in texts from the geography domain and *pământ* "earth" in the others). Additional difficulties arose in distinguishing adverbs from nouns in temporal expressions, where the syntactic context was insufficient for reliable automatic disambiguation.

Further challenges included homonymy within verb forms across tenses (present of the indicative mood and simple past of the indicative mood), ambiguous clitic pronouns with identical surface forms (e.g., *le* can be either dative or accusative), and nouns sharing forms between singular oblique and plural direct cases (e.g., *fete*). Abbreviations were frequently misclassified as proper nouns, requiring manual correction, while lemmatization errors affected neuter and irregular plural forms. These issues highlight the limitations of automatic morphological processing for Romanian and underscore the necessity of systematic human validation.

[4]The form *îs* "is" is regional nowadays.

3.2 Semantic Annotation

Two types of semantic information are annotated in ELEXIS-Ro: multiword expressions and named entities.

3.2.1 Multiword expressions

Multiword expressions (MWEs) are combinations of at least two lexicalized components that form a weakly connected structure and exhibit orthographic, morphological, syntactic, or semantic idiosyncrasies relative to general grammar (Baldwin and Kim, 2010), following the PARSEME guidelines[5]. The corpus covers a broad typology of MWEs:

- Verbal
  - Verbal Idioms (VID): *avea în vizor* (‘have in sight’ “keep in view, target”), *da nas în nas* (‘give nose in nose’ “run into, bump into”)
  - Light Verb Constructions (LVC)
    - full: *avea întâlnire* (‘have meeting’ “have a meeting”), *da citire* (‘give reading’ “read”)
    - cause: *da asigurare* (‘give assurance’ “assure”), *pune în aplicare* (‘put in application’ “implement, carry out”)
  - Reflexive Verbs (IRV): *-și permite* (‘3SG.DAT.REFL allow’ “afford”), *se abține* (“refrain”)
  - Inherently Adpositional Verbs (IAV): *se baza pe* (‘3SG.ACC.REFL base on’ “rely on, be based on”), *depinde de* (“depend on”)
- Nominal
  - Nominal Idioms (NID): *act de identitate* (‘act of identity’ “identity card”), *an de grație* (‘year of grace’ “year of Our Lord, A.D.”)
  - Pronominal Idioms (PronID): *câte ceva* (“something”), *Domnia Sa* (‘His/Her Lordship’ “His Excellency”)
  - Deverbal Nominal MWEs (NV): *punere în funcțiune* (‘putting in function’ “commissioning, putting into operation”), *băgare de seamă* (‘putting of notice’ “attention, observation”)
- Modifier
  - Adjectival Idioms (AdjID): *cu o falcă în cer și cu una în pământ* (‘with one jaw in sky and one in earth’ “very furious”), *cu cântec* (‘with song’ “with flair, extravagantly”)
  - Adverbial Idioms (AdvID): *așa cum se cuvine* (‘so as 3SG.ACC.REFL befi’ “properly, as it should be”), *ca oamenii* (‘like people.THE.PL’ “like decent people, properly”)
  - Deverbal adjectival / adverbial MWEs (AV): *cu luare aminte* (‘with taking notice’ “attentively”), *luat în seamă* (‘taken in notice’ “taken into account, considered”)
- Functional
  - Determiner Idioms (DetID): *picior de* (‘foot of’ “not a single”), *ca atare* (“as such”)
  - Adposition Idioms (AdpID): *de dragul* (‘of dear.DEF.SG.M’ “for the sake of”), *cu excepția* (‘with the exception’ “except for”)
  - Conjunction Idioms (ConjID): *astfel încât* (“so that”), *pe motiv că* (‘on reason that’ “because, on the grounds that”)
  - Interjection Idioms (IntjID): *ce folos* (‘what use’ “what’s the point?”), *nici vorbă* (‘no word’ “no way, not a chance”)

Automatic MWE identification was performed using a lexicon of MWEs derived from the validated PARSEME-Ro corpus (Barbu Mititelu et al., 2025), yielding 7,268 unique occurrences mapped to PARSEME labels and matched against ELEXIS-Ro. The matching procedure allowed for discontinuous realizations by permitting up to two intervening tokens between components, balancing recall and noise in the detection of long-distance dependencies.

Automatic identification procedures inevitably introduce both false positives and false negatives, due to the structural variability and context-dependent interpretation of MWEs. In line with established methodologies, we performed a systematic validation of the automatically extracted and annotated MWEs, focusing on verifying their boundaries, eliminating spurious candidates, and ensuring the correct assignment of MWE

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

F1	A1	A2	A3
A1	1	0.7681	0.7587
A2	0.7681	1	0.8547
A3	0.7587	0.8547	1

Table 1: Inter-annotator agreement

types. Particular attention was paid to ambiguous cases, homonymous expressions, and constructions whose MWE status depends on contextual interpretation.

The validation process involved three trained annotators and followed a cross-validation protocol inspired by previous annotation campaigns. Each candidate MWE was independently assessed by two annotators, applying the PARSEME decision tests and annotation guidelines. The corpus was split in five files, each of 500 sentences (except the fifth one). Three annotators took care of the validation and correction of their MWE annotation, the pairing among them being different from file to file: file 1 was annotated by annotators A1 and A2, file 2 by annotators A2 and A3, file 3 by annotators A1 and A3, file 4 by annotators A1 and A2, file 5 by annotators A2 and A3.

The inter-annotator agreement for each file is shown in Table 1. We adopted a pairwise overlapping validation design, where each sentence was validated and corrected by two out of three annotators, ensuring balanced coverage of all annotator pairs while controlling annotation cost. Inter-annotator agreement was computed following the PARSEME shared task methodology (Ramisch et al., 2018), where agreement is defined as an F1-score over MWE annotations (Fspan), based on exact span and category matching [6]. Validation results coming from the three annotators were considered independent annotation sets, and pairwise F1 scores were computed for all annotator pairs (see Table 1). The overall agreement score (0.7938) was obtained by averaging these pairwise F1 values.

3.2.2 Named entities

Named entities (NEs) of three types were annotated in the corpus: person (PER), organization (ORG) and location (LOC). Just like the other types of annotation, this one was also done automatically,

[6]Fleiss' kappa was not used because it assumes independent categorical decisions and does not account for span-based, possibly overlapping MWE annotations.

PoS	Number		
NOUN	8413	ADV	1337
ADP	4924	PRON	1187
PUNCT	4110	NUM	1178
VERB	3109	CCONJ	1019
ADJ	2653	PART	484
AUX	2372	SCONJ	234
DET	2322	X	222
PROPN	1750	INTJ	1
		SYM	1
	TOTAL	**35316**	

Table 2: PoSes distribution in ELEXIS-Ro

using three lists of NEs[7]. The automatic annotation was checked by two annotators and corrections were made when necessary (skipped NEs, strings incorrectly classified as NEs, incorrect length of NEs, incorrect label). When the two annotators disagreed in their evaluation, a third, specialist, one made the final decision.

4 Statistics on ELEXIS-Ro

4.1 General statistics

As for all the other languages it contains, the corpus has 2024 sentences. In Romanian these sentences contain 35,316 words, with an average sentence length of 17.45.

4.2 Statistics on PoSes

The distribution of tokens according to the part-of-speech can be seen in Table 2. All parts of speech are present in the corpus, with nouns dominating, which is indicative of an information-dense corpus.

4.3 Statistics on MWEs

Overall, 3,018 MWEs were automatically identified in the Romanian corpus, and their number of occurrences and relative frequencies in the ELEXIS-Ro corpus are rendered in column two and three of Table 4. For their comparison with the MWEs in the PARSEME-Ro corpus, see Section 5 below.

4.4 Statistics on NEs

The distribution of the three types of NEs in the corpus is presented in Table 3. This is based on the automatic identification (the middle column),

[7]These lists are available at https://www.racai.ro/p/reterom/rezultate.html, last accessed April 2026

Type of NE	#	Final #	% skipped
PER	277	516	53
ORG	220	321	68
LOC	404	754	53
TOTAL	**901**	**1591**	

Table 3: NEs distribution in ELEXIS-Ro: column two contains the numbers after automatic annotation, while the third column renders the number after the manual validation, correction and clarification of disagreements.

with the results of the manual validation in the third column (see Section 6). We notice that about half of the NEs were not recognized automatically, with the distribution per type shown in the last column.

5 Discussion of results

MWEs, as recurrent lexical or phrasal patterns, often exhibit variation in frequency, syntactic structure, and semantic category depending on the register and domain of the texts in which they occur. We are going to discuss now the quantitative differences between the MWEs occurring in the ELEXIS-Ro and in the PARSEME-Ro corpora. The former is primarily composed of encyclopedic and informative texts, while the latter consists exclusively of journalistic texts, thus they provide contrasting linguistic environments. This contrast makes it possible to examine not only absolute frequencies, but also the relative distribution of MWE types, offering a basis for interpreting the statistical patterns discussed below.

It should also be noted that the two corpora differ considerably in terms of the total number of annotated MWEs, with 3,018 instances in ELEXIS and 64,601 in PARSEME-Ro. This difference is due to variations in corpus size and composition. In order to ensure comparability, the analysis is therefore based on relative frequencies, which provide a more appropriate basis for examining the distribution of MWE types across the two datasets. Accordingly, the discussion focuses on proportional patterns rather than absolute counts.

The quantitative comparison reveals both convergences and divergences in the distribution of MWE types across the two corpora. At the top of the distribution, AdpID expressions show a remarkably similar proportion in both datasets, accounting for 28.60% in ELEXIS and 28.48% in PARSEME-Ro. This consistency suggests that this category is relatively stable across genres and less sensitive to variation in text type. A similar, though slightly less precise, pattern can be observed for ConjID expressions (6.83% in ELEXIS vs. 6.51% in PARSEME-Ro) and, to a lesser extent, DetID, which displays nearly identical proportions in both corpora. These categories appear to form a core layer of MWEs whose distribution remains largely unaffected by domain-specific factors.

In contrast, several categories exhibit substantial variation. AdvID expressions are considerably more frequent in ELEXIS (28.00%) than in PARSEME-Ro (14.50%), pointing to a stronger presence of adverbial multiword units in encyclopedic and informative texts, where such structures contribute to descriptive precision and elaboration.

Conversely, PARSEME-Ro shows higher proportions of verb-related MWEs and structurally more complex expressions. Categories such as IAV (12.14% vs. 8.35%), IRV (9.98% vs. 5.86%), and especially NID (11.93% vs. 3.08%) are much more prominent in the journalistic corpus. This pattern suggests a preference for more dynamic, predicate-centered constructions, as well as noun-based MWEs that may encode salient or context-dependent information.

A similar tendency can be observed for LVC.full and LVC.cause, which are more frequent in PARSEME-Ro, although they remain relatively marginal overall. These constructions, often associated with more complex predicate formation, further point to a greater use of analytically structured verbal expressions in journalistic texts.

Lower-frequency categories (e.g., AV.VID, NV.VID, NV.LVC.full) remain marginal in both corpora, and their small proportions make it difficult to draw strong genre-based conclusions. However, even in these cases, PARSEME-Ro tends to display slightly higher values and more diverse types, reinforcing the general pattern of greater structural diversity.

Overall, the comparison highlights a distinction between more stable, function-word-based MWE categories and those that are more sensitive to genre. While the former remain consistent across corpora, the latter vary more noticeably, reflecting the role of textual domain in shaping both the frequency and typology of MWEs.

6 Future Work

On the one hand, validation and correction of the automatic annotation is the very next step. This will

MWE type	# in ELEXIS-Ro	% in ELEXIS-Ro	# in PARSEME-Ro	% in PARSEME-Ro
AdpID	863	28.60	18401	28.48
AdvID	845	28.00	9366	14.50
AdjID	347	11.50	4496	6.96
IAV	252	8.35	7840	12.14
ConjID	206	6.83	4206	6.51
IRV	177	5.86	6449	9.98
NID	93	3.08	7709	11.93
VID	79	2.62	2573	3.98
PronID	73	2.42	745	1.15
AV.IAV	30	0.99	779	1.21
LVC.full	18	0.60	934	1.45
DetID	17	0.56	387	0.60
AV.VID	6	0.20	28	0.04
NV.LVC.cause	4	0.13	76	0.12
IntjID	4	0.13	59	0.09
LVC.cause	2	0.07	311	0.48
NV.VID	1	0.03	106	0.16
NV.LVC.full	1	0.03	67	0.10
NV.IAV	–	–	54	0.08
AV.LVC.cause	–	–	11	0.02
AV.LVC.full	–	–	4	0.01
NV.IRV	–	–	1	0.00
TOTAL	**3018**		**64601**	

Table 4: Comparative statistics between MWEs occurring in ELEXIS-Ro and in PARSEME-Ro.

target syntactic parsing correctness, in line with the UD principles of annotation, deciding upon the cases of disagreement between the annotators of the MWEs.

On the other hand, the corpus will be semantically disambiguated, i.e. each content word will be assigned a sense, based on the Romanian WordNet (Tufiș and Mititelu, 2015).

Once the corpus annotation is validated and corrected it can serve as a reliable source of linguistic information for comparative studies.

7 Conclusions

In this paper, we have presented the extension of the ELEXIS parallel corpus through the integration of Romanian, following a hybrid workflow that combines automatic processing with systematic manual validation. The pipeline included translation from English, tokenisation, morpho-syntactic annotation with UDPipe, and semantic enrichment through the identification and classification of MWEs and NEs. The results confirm that while automatic tools provide a solid baseline for large-scale corpus construction, they consistently produce errors in contexts involving ambiguity, homonymy, and language-specific structures, making manual correction an essential component for ensuring annotation quality.

The analysis of annotation errors highlights several recurrent challenges specific to Romanian, including part-of-speech ambiguity, lemmatisation errors, and difficulties in handling clitics, compounds, and NEs. Similarly, the automatic identification of MWEs, based on the PARSEME guidelines, proved effective, but required careful post-editing for the correct span identification and type of MWE labelling. The double validation process contributed to improving consistency and reliability across annotation layers, reinforcing the importance of human expertise in multilingual resource development.

Overall, the Romanian extension enhances the ELEXIS corpus both quantitatively and qualitatively, expanding its coverage to a less-resourced language and increasing its value for cross-linguistic research. The resulting dataset, enriched with detailed morpho-syntactic and semantic annotation, provides a robust resource for studies in lexicography, translation, and multilingual natural language processing.

8 Corpus Availability

ELEXIS-Ro has been released in the 2.0 version of the ELEXIS corpus (Čibej et al., 2026).

9 Acknowledgment

This work was supported by a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV. This work also received support from the CA21167 COST action UniDive, funded by the European Union via the COST (European Cooperation in Science and Technology).

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton, USA.

Verginica Barbu Mititelu. 2018. Modern syntactic analysis of Romanian. In *Clasic și modern în cercetarea filologică românească actuală*, pages 67–78, Iași, Romania.

Verginica Barbu Mititelu, Ioana-Maria Biolan, Ioana Buhnila, Mihaela Cristescu, Elena Irimia, Catalin Mihaila, Isabella Sinca, Amalia Todirascu, and Carmen Vasile. 2025. Enhancing the parseme-ro corpus with nominal, modifier and functional multiword expressions. In *Proceedings of the 20th International Conferene on Linguistic Resources and Tools for Natural Language Processing*, pages 113–125.

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, Sia Kolkovska, Kadri Muischnek, Kertu Saul, Karoliina Jõgi, Mija Bon, Ranka Stanković, Cvetana Krstev, Aleksandra Marković, Milica Ikonić Nešić, Voula Giouli, Eri Papanikolaou, Irina Lobzhanidze, Verginica Barbu Mititelu, Simina Popa, Lea Cristiana, Mihaila Catalin, Elena Irimia, Ana Ostroški Anić, Siniša Runjaić, Robert Sviben, Martina Pavić, Ivana Filipović Petrović, Bartłomiej Alberski, Vladimir Cvetkoski, Olha Kanishcheva, and Rusudan Makhachashvili. 2026. Parallel sense-annotated corpus ELEXIS-WSD 2.0. Slovenian language resource repository CLARIN.SI.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Cvetana Krstev, Ranka Stanković, Aleksandra M. Marković, and Teodora Sofija Mihajlov. 2024. Towards the semantic annotation of SR-ELEXIS corpus: Insights into multiword expressions and named entities. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 106–114, Torino, Italia. ELRA and ICCL.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, and Tina Munda. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. *Proceedings of Electronic Lexicography in the 21st Century Conference*, (2021):377–395. UIDB/03213/2020 UIDP/03213/2020; eLex 2021, 7th biennial conference on electronic lexicography, eLex 2021 ; Conference date: 05-07-2021 Through 07-07-2021.

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

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.

Agata Savary, 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, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

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.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Dan Tufiș and Verginica Barbu Mititelu. 2015. *The Lexical Ontology for Romanian*, pages 491–504. Springer International Publishing, Cham.