ELEXIS-Ro – Adding Romanian to the ELEXIS Corpus

Verginica Barbu Mititelu
Romanian Academy, Research Institute for Artificial Intelligence
Bucharest, Romania
vergi@racai.ro

Elena Irimia
Romanian Academy, Research Institute for Artificial Intelligence
Bucharest, Romania
elena@racai.ro

Cătălin Mihăilă
Romanian Academy, Research Institute for Artificial Intelligence
Bucharest, Romania
catalin@racai.ro

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

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

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

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

The paper presents, against the landscape of corpora dedicated to or also containing the Romanian language, the extension of the ELEXIS corpus with yet another language, Romanian, which both enhances the value of the parallel corpus and offers Romanian new opportunities for comparative and contrastive research within the Romance language family, as well as with the Balkan languages.

The Romanian component has been integrated primarily through automatic methods, followed by systematic manual validation to ensure annotation quality and cross-linguistic consistency. The levels of processing and annotation are: automatic translation of the English corpus into Romanian, automatic tokenization of the sentences, automatic morpho-syntactic annotation and semantic annotation (i.e., annotation with multiword expressions).

The annotation frameworks are widely used, multilingual ones: for morphology and syntax we used Universal Dependencies, while for multiword expressions, we made use of PARSEME guidelines.

Keywords: parallel corpus, Romanian, morpho-syntactic annotation, multiword expressions

https://doi.org/10.47810/JCLIB.2.2026.02

1. Introduction

Parallel corpora, defined as collections of texts aligned across two or more languages, constitute essential linguistic resources for multilingual research and language technology development. By providing direct correspondences between original texts and their translations, parallel corpora enable systematic investigation of how meaning, structure, and discourse features are transferred across languages. These resources facilitate the identification of translation patterns, equivalence relations, and cross-linguistic variation, offering valuable insights into both universal and language-specific phenomena. When translations included in such corpora are carefully produced and validated – either by professional translators or through rigorous quality control procedures – the resulting datasets become reliable repositories of linguistic and conceptual knowledge.

Despite the rapid advancement and widespread adoption of Large Language Models (LLMs) in recent years, parallel corpora remain indispensable. While LLMs demonstrate remarkable capabilities in multilingual understanding and generation, their performance still depends heavily on the availability of high-quality multilingual training data. Parallel corpora provide structured and verifiable bilingual or multilingual alignments that can be used to evaluate, fine-tune, and interpret these models. Furthermore, unlike automatically generated multilingual data, curated parallel corpora offer transparent mappings between languages, allowing researchers to assess the fidelity of meaning transfer and to analyze linguistic phenomena with a higher degree of confidence. In this sense, parallel corpora continue to play a crucial role in grounding data-driven approaches and ensuring that multilingual language technologies remain interpretable and reliable.

By incorporating low-resourced languages into parallel datasets, researchers not only preserve linguistic diversity but also enable cross-lingual transfer learning, domain adaptation, and the development of multilingual systems that perform more equitably across languages. Such corpora can support tasks including machine translation, cross-lingual information retrieval, terminology extraction, and linguistic typology, while also contributing to language preservation and revitalization efforts.

Moreover, the usefulness of a parallel corpus increases substantially with the addition of multiple levels of linguistic annotation. Beyond sentence-level alignment, annotations may include morphological, syntactic, semantic, discourse, and pragmatic information, as well as terminology, named entities, and domain-specific labels. Multi-layer annotation enables researchers to investigate correspondences across linguistic levels, revealing how grammatical structures, semantic roles, discourse relations, and stylistic features are translated or adapted across languages. These enriched corpora facilitate more sophisticated analyses and support downstream applications such as supervised machine translation, cross-lingual parsing, multilingual terminology management, and knowledge extraction.

In this context, richly annotated multilingual parallel corpora represent high-value resources for both theoretical and applied research. They not only support linguistic analysis and comparative studies, but also serve as valuable datasets for training, evaluating, and improving multilingual language technologies. As research increasingly moves toward multilingual and inclusive approaches, the development of high-quality, multi-layered parallel

corpora – especially those including low-resourced languages – remains a priority for advancing both linguistic understanding and language processing.

Against this background, we present the extension of the ELEXIS parallel corpus (Martelli et al. 2021) through the addition of Romanian. This inclusion represents a significant enhancement of the resource, both in terms of linguistic coverage and research potential. With the integration of Romanian, the corpus now includes a fourth Romance language, alongside the existing Romance-language components. This enrichment not only increases the multilingual scope of the corpus, but also opens new opportunities for comparative and contrastive research within the Romance language family. Researchers will be able to investigate similarities and differences in lexical choices, syntactic structures, semantic mappings, and discourse strategies across closely related languages, thereby contributing to a deeper understanding of language variation and evolution within the Romance domain.

Furthermore, the addition of Romanian is particularly valuable from a typological and areal linguistic perspective. Romanian occupies a distinctive position within the Romance language family due to its geographical location in Eastern Europe and its long-standing contact with Slavic and other Balkan languages. As a result, Romanian exhibits features that differentiate it from Western Romance languages, including structural and lexical characteristics influenced by language contact. The availability of Romanian within the ELEXIS parallel corpus therefore enables researchers to explore not only genealogical relationships within the Romance family, but also contact-induced phenomena and convergence patterns.

In this respect, the simultaneous presence of Bulgarian in the corpus further enhances its research potential. Romanian and Bulgarian are both part of the Balkan linguistic area (the Balkan Sprachbund), where languages from different families share common structural features due to prolonged contact and mutual influence. The inclusion of both languages within the same parallel corpus makes it possible to conduct systematic comparisons of Balkan-specific phenomena, such as definiteness marking, analytic constructions, clitic doubling, and other shared morphosyntactic traits. Consequently, the extended corpus supports investigations at multiple levels: genealogical comparisons among Romance languages, typological comparisons across language families, and areal studies within the Balkan linguistic context.

Overall, the addition of Romanian strengthens the ELEXIS parallel corpus as a multilingual and multi-purpose resource, broadening its applicability for contrastive linguistics, translation studies, typology, and multilingual Natural Language Processing (NLP).

The ELEXIS corpus is a manually annotated multilingual parallel dataset designed to support both lexicographic research and Word Sense Disambiguation (WSD). The project was first introduced in 2021 and initially developed for ten European languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish. The dataset consists of parallel sentences extracted from WikiMatrix, an open-access collection of aligned sentences derived from Wikipedia. These sentences were first automatically selected and subsequently manually validated to ensure linguistic quality and semantic clarity. Additional selection criteria were applied, including a minimum length of five words per sentence and the presence of at least two polysemous lexical items, resulting in a carefully curated corpus comprising 2,024 sentences.

The corpus is enriched with five layers of linguistic annotation: tokenisation, sub-tokenisation, lemmatisation, part-of-speech tagging, and WSD. These annotation layers provide a detailed linguistic representation that supports both fine-grained semantic analysis and cross-linguistic comparison. In particular, the WSD layer plays a central role in the project, as it enables the identification and alignment of word senses across multiple languages. This multilingual semantic annotation supports lexicographic work by refining sense inventories, identifying gaps or inconsistencies in existing lexical resources, and facilitating the development of more precise and comprehensive dictionaries. At the same time, the dataset provides a valuable benchmark for evaluating and improving WSD systems, particularly in multilingual and cross-lingual contexts.

The primary objective of the ELEXIS project is to complete and enhance the semantic annotation across all participating languages while simultaneously improving and expanding existing sense inventories. By offering consistent semantic annotation across multiple languages, the corpus establishes a new reference point for multilingual WSD systems and promotes interoperability among lexical resources. Moreover, the dataset contributes to advancing multilingual NLP by providing high-quality, manually curated data that can be used for training, evaluation, and comparative analysis. This is particularly important given that current NLP resources and models are still heavily biased toward English, resulting in uneven technological development across languages.

In this context, the addition of Romanian represents a valuable and timely extension of the corpus. Romanian remains relatively under-resourced in comparison to major European languages, particularly in the area of semantically annotated multilingual datasets. Including Romanian in the ELEXIS corpus would expand linguistic coverage, enhance the representation of Eastern European languages, and support cross-lingual semantic research involving both Romance and Balkan languages. Furthermore, the availability of Romanian data in a richly annotated multilingual corpus would facilitate the development and evaluation of NLP tools for Romanian, including WSD systems, machine translation, semantic parsing, and lexicographic applications.

The paper is structured as follows: we present the landscape towards which we paint the development of this new corpus, part of a parallel one (Section 2): first, we outline the results of the work in corpora development for Romanian (Subsection 2.1) and then we present other endeavours for extending the ELEXIS corpus (Subsection 2.2). We continue with the steps taken in the development of the ELEXIS-Ro, the Romanian component of the ELEXIS corpus (Section 3): the translation of the English component into Romanian (Subsection 3.1) and then each of the annotation phase in turn: tokenization (Subsection 3.2), morpho-syntactic annotation (Subsection 3.3) and the semantic annotation of the multiword expressions (Subsection 3.4). Concluding the paper, we also announce the future work, i.e., the new annotation levels.

2. Related work

2.1 Romanian corpora

Romanian language corpora have been developed both as components of multilingual resources and as independent monolingual collections, each serving different research

purposes and covering a variety of linguistic domains. These corpora constitute essential resources for NLP, computational linguistics, and corpus-based linguistic studies, providing structured linguistic data for tasks such as machine translation, terminology extraction, and language modeling.

Within the category of multilingual corpora that include Romanian, several significant resources can be identified. One of the most important is the RO-JRC-Acquis (Ceauşu 2008), which contains over 30 million words. This corpus is derived from the European Commission Joint Research Centre's legislative documents and reflects predominantly the legal domain. Due to its parallel structure and terminological consistency, the corpus is particularly valuable for legal language processing and multilingual alignment studies.

Another important multilingual resource is the EUR-Lex Judgements Corpus (Baisa et al. 2016), which contains more than 17 million words. This corpus consists of judicial decisions from European Union institutions and, similarly to RO-JRC-Acquis, reflects the legal domain. Its structured format and domain specificity make it suitable for legal text mining, cross-lingual legal research, and computational legal linguistics.

The EUROPARL Corpus (Koehn 2005) represents another widely used multilingual dataset, containing nearly 10 million Romanian words extracted from the proceedings of the European Parliament. This corpus is particularly valuable for studying formal spoken language, parliamentary discourse, and translation patterns across European languages.

A larger multilingual resource that includes Romanian is OPUS (Tiedemann 2012), which contains approximately 300 million Romanian words. OPUS aggregates multiple parallel corpora collected from various sources, including legislative, administrative, and institutional texts. Due to its size and diversity, OPUS is frequently used for training machine translation systems and for large-scale multilingual language modeling.

In addition to multilingual corpora, several monolingual Romanian corpora have also been developed. Among these, the RoCo-news Corpus (Tufiş and Irimia 2006) contains approximately 7 million tokens and primarily reflects journalistic language. This corpus is particularly useful for studies focusing on contemporary Romanian media discourse and stylistic analysis.

Another significant monolingual resource is the RoWaC (Macoveiciuc and Kilgarriff 2010), which contains nearly 45 million words collected from web sources. RoWaC captures a wide range of informal and semi-formal language usage, making it valuable for studying contemporary linguistic variation, web-based discourse, and emerging vocabulary.

The ROMBAC (Ion et al. 2012) represents a carefully designed balanced corpus containing approximately 36 million words. The corpus is distributed relatively evenly across five domains: journalistic texts, pharmaceutical and medical texts, legal documents, literary history, and fiction. Due to this balanced composition, ROMBAC is particularly suitable for general-purpose linguistic research and domain comparison studies and it was at the core of the design and development of the next major Romanian corpus, CoRoLa.

The culmination of these efforts of corpus creation was the development of the reference corpus designed to reflect contemporary written and spoken Romanian language usage, CoRoLa (Barbu Mititelu, Tufiş, and Irimia 2018), which represented a priority research initiative of the Romanian Academy. This project was coordinated by two of its specialized

institutes, namely the Research Institute for Artificial Intelligence in Bucharest and the Institute for Computer Science in Iași. Despite this institutional coordination, the creation of the corpus evolved into a nationwide collaborative effort. In addition to this, the project also incorporated an important international component. Through the DRuKoLA Project, funded by the Alexander von Humboldt Foundation, the corpus – CoRoLa – benefited from a robust technological infrastructure for indexing and querying its content (Diewald et al. 2016).

At the time of its development, CoRoLa was conceived as a multipurpose resource designed to meet the needs of several user communities. For linguists, the corpus provides empirical evidence for a wide range of linguistic phenomena, allowing researchers to investigate frequency patterns, stylistic variation, domain-specific language use, and other characteristics of contemporary Romanian. The availability of large-scale, naturally occurring textual data enables both qualitative and quantitative linguistic analyses. For language engineers and researchers in NLP, CoRoLa serves as a valuable resource for training computational models, including word embeddings and other statistical representations of language. These representations, extracted from the corpus, support a wide range of applications such as automatic text classification, information retrieval, machine translation, and semantic analysis. At the same time, the corpus is also intended for use by the general public. Users can consult the corpus to identify authentic language usage examples, explore word meanings in context, and examine collocational patterns. This functionality makes CoRoLa a useful tool not only for academic research, but also for language education, lexicography, and professional writing. More recently, international collaborative projects have provided new opportunities to resume and expand the corpus. Particular attention has been paid to ensuring the authenticity and reliability of the data, especially in the context of the growing prevalence of AI-generated text. To minimize the inclusion of artificially generated content, web crawling activities were restricted to materials published prior to 2023. This precautionary measure aims to preserve the corpus as a reliable representation of naturally occurring contemporary Romanian language usage (Irimia et al. 2026).

Together, these corpora provide extensive coverage of Romanian language usage across multiple domains and registers, forming a solid foundation for both linguistic analysis and NLP applications.

2.2 Extension of ELEXIS with other languages

The ELEXIS project[1] has led to the development of a parallel sense-annotated corpus across ten languages (see above, Section 1). During the UniDive COST Action (Savary et al. 2024), the corpus was extended with several new languages, as shown below.

The Danish component, DA-ELEXIS, (Pedersen et al. 2023) is one of the largest sense-annotated corpora available for Danish and constitutes an important component of the broader ELEXIS multilingual infrastructure. The corpus was designed with multiple annotation layers, including tokenisation, sub-tokenisation, lemmatisation, and part-of-speech

1 https://project.elex.is/

(POS) tagging. These initial layers were automatically generated following the Universal Dependencies (de Marneffe et al. 2021) guidelines and subsequently manually verified to ensure annotation accuracy and consistency.

The Danish corpus contains 32,524 tokens, with sense annotations provided for all content words. These annotations include 7,322 nouns, 3,099 verbs, 2,626 adjectives, and 1,677 adverbs[2]. The annotation workflow was supported by a dedicated annotation tool that guided annotators through each token in its sentential context. For each token, the system presented all available senses, allowing annotators to select the most appropriate meaning based on contextual interpretation. This semi-automated annotation process facilitated consistency while maintaining human oversight.

The primary sense inventory used for senses annotation was DanNet, which covers approximately 70,000 Danish lemmas. In cases where compound words were not present in the sense repository, the DA-ELEXIS corpus applied compound splitting into constituent lemmas that could be found therein. This strategy enabled semantic tagging of otherwise unavailable lexical items. However, compounds already present in DanNet were preserved as single units.

Multiword expressions (MWEs) were encoded using a single semantic label. When MWEs appeared discontinuously in the corpus, the entire expression was subsequently identified and linked to lexical resources to ensure consistent annotation.

To evaluate annotation reliability, slightly more than 5% of the corpus (108 sentences) was triple-annotated. Inter-annotator agreement was measured using Cohen's kappa, resulting in an average score of 0.68. After clarifying annotation guidelines and resolving inconsistencies, agreement improved to 0.78 in later stages of the project, indicating satisfactory annotation reliability for semantic tagging tasks.

Krstev et al. (2024) present their work on the Serbian extension of the ELEXIS. The project focuses primarily on semantic annotation and the creation of lexical sense repositories, while also exploring multilingual correspondences across all ten languages included in ELEXIS, particularly for MWEs and named entities (NEs). The sentences from WikiMatrix were translated from English into Serbian using Google Translate, followed by manual verification. Special attention was given to incorrectly translated MWEs, morphological inconsistencies, and phonetic transcriptions of NEs. Tokenisation, lemmatisation, and POS tagging were initially performed automatically and subsequently manually validated to ensure annotation quality.

Because the number of MWEs and NEs varies across languages, expressions from all ten languages were automatically translated into Serbian as phrases rather than word-by-word equivalents. This process revealed structural differences between languages, including instances where MWEs in other languages correspond to single words in Serbian and vice versa. The annotation process produced 529 non-verbal MWE occurrences (i.e. 339 unique ones): 351 nominal MWEs, 133 proper nouns, 44 adverbial expressions, and one adjectival expression. Additionally, 230 occurrences of verbal MWEs were identified, representing 98 distinct expressions. No adjectival or verbal similes were found in this dataset.

2 These numbers can be compared with the corresponding ones in Romanian from Table 2.

Named Entity Recognition (NER) tools identified multiword NEs such as organizations and geopolitical names, which were recognized both by lexical dictionaries and by automatic NER systems. The applied systems are designed to identify new entities when compared with datasets in other languages or following manual validation, allowing the Serbian repository to expand iteratively.

The primary lexical resource used for Serbian word sense annotation was the Serbian WordNet (SrpWN) (Krstev et al. 2004). The English ELEXIS sense inventory is based on the Princeton WordNet (PWN) (Miller 1995; Fellbaum 1998). Because the dataset did not include the WordNet interlingual index, alignment between the two datasets was achieved by comparing definitions. Synonym lists from PWN were automatically translated into Serbian using both Google API and OpenAI technologies. Additional lexical resources from previous research projects were also incorporated to expand the Serbian subset. The results indicated that approximately 39% MWEs from the Serbian ELEXIS dataset were found in the SrpWN. Some entities were present in multiple synsets, while in other cases Serbian used single lexical items where other languages employed MWEs. These findings highlight both cross-linguistic variation and the challenges of multilingual semantic alignment.

3. Methodology

In this section we focus on the work done for adding Romanian to the ELEXIS corpus. At the moment of this writing, the following steps have already been taken: translation of the English corpus, tokenization of the Romanian corpus, morpho-syntactic annotation, semantic annotation with types of MWEs. Each step is described in what follows.

In accordance with the methodology adopted in the ELEXIS parallel corpus, the translation and annotation of the Romanian component are initially performed automatically using a range of specialized tools for machine translation, tokenization, lemmatization, and part-of-speech tagging. These automatically generated annotations are subsequently manually revised, corrected, and validated in order to ensure linguistic accuracy, consistency across languages, and compatibility with the multilingual sense-annotation framework. This combined automatic–manual workflow enables efficient corpus expansion while maintaining high-quality annotation standards required for cross-linguistic semantic analysis and lexicographic applications.

3.1 Translation

Automatic translation from English into Romanian was done using Google Translate. During manual inspection, 1,437 sentences (out of 2024) required no manual interventions, thus were considered correct translations. This means the rest 30% of the sentences were manually corrected.

Machine translation (MT) errors cluster around a limited number of recurring patterns where human intervention becomes systematic and predictable. The first and most frequent category concerns lexical and register-related errors, namely the selection of formally correct equivalents that are nevertheless contextually or stylistically inappropriate:

- MT systems often produce "acceptable" but imprecise or stylistically neutral synonyms, such as *se prezintă* instead of *apare* for the meaning "appears", *au vizitat premiera* instead of *au participat la premiera* "attended (the premiere)", *lucrarea sa ulterioară* instead of *opera sa ulterioară* referring to "his later work", or *rezolvarea completă* instead of *vindecarea completă* for "complete recovery".
- The same phenomenon occurs in terminological contexts, where *cursuri de master* is used instead of *masterclass-uri*, *raze X* instead of *radiografii* for the translation of "X-rays" as a medical investigation procedure, or *grupare funcțională* is used instead of *grupă funcțională*. In such cases, the human translator acts as a filter of contextual appropriateness, selecting forms that correspond to the appropriate register – technical, academic, or colloquial – and to natural Romanian usage, without altering the underlying meaning.
- This category also includes unnatural collocations and literal translations, where MT reproduces source language structures: e.g., *a acordat un rating mare* "gave a high rating", or *împărtășește o graniță* "shares a border", which are subsequently normalized by human intervention into *a lăudat* "praised" and, respectively, *are o graniță comună* "has a common border", as well as non-idiomatic formulations such as *orașul a fost stabilit* instead of *orașul a fost întemeiat* for the English "the city was established", or *se bazează pe* instead of *se inspiră din* for the English "is inspired by".

The second major category involves morphosyntactic and structural errors, where MT transfers patterns from the source language or produces incorrect agreement and syntactic relations. These include:

- agreement errors: *Majoritatea populației este catolic* instead of *catolică* (i.e., MT used a masculine instead of a feminine adjective), *au devenit membră* instead of *membre* (singular instead of pluaral),
- incorrect determiner: *oricui pilot* instead of *oricărui pilot* (*oricui* is not a determiner, unlike *oricărui*),
- incomplete or ambiguous constructions: *Dacă din cauza unui virus…* instead of *Dacă este cauzată de un virus…*,
- issues of word order and fluency also arise when MT preserves source-language structure, requiring reordering for clarity: *vocea lui Nick i-a plăcut* instead of *i-a plăcut vocea lui Nick*.

An important subcategory concerns the handling of NEs and linguistic conventions, including capitalization (*primul război mondial* instead of *Primul Război Mondial*), adaptation of proper names and titles, and localization of acronyms and institutional names (*Regional Internet Registry* translated as *Registru Regional al Internetului*).

Overall, the essential difference lies in the fact that MT typically delivers translations that are semantically correct, but structurally dependent on the source language, whereas

human translators intervene along lexical, syntactic, and stylistic dimensions to produce coherent, natural texts fully aligned with the norms of the target language.

3.2 Tokenisation

Automatic tokenisation was carried out with UDPipe[3] (Straka, Hajic, and Straková 2016) using a Romanian model[4] based on the Romanian Reference Treebank[5] (RRT) (Barbu Mititelu and Irimia 2016). The resulting tokenised sentences were then distributed for manual correction in a shared Google Sheets document, accompanied by a concise set of guidelines outlining annotation instructions common to all participating languages.

The guidelines emphasized the importance of accurate tokenisation as a foundation for all subsequent layers of annotation, including lemmatization, POS tagging, UD parsing, MWE annotation, NE annotation, and WSD. At this stage, each language team must decide how to treat multiword compounds at the WSD annotation level, more specifically, whether to WSD-annotate the individual tokens within a compound. If disambiguation of the elements is decided, compounds should be treated as multiword tokens and tokenised accordingly. In the case of Romanian, we identified only one situation requiring such token-level disambiguation, namely ad-hoc compounds (see Table 1).

Tokenisation correction is performed by adding or deleting lines in the Google Spreadsheet file, with due attention to the "SpaceAfter=No" attribute in the last column (see last column of Table 1), which indicates that a token is not followed by a whitespace in the original sentence. This attribute is essential, as it allows the reconstruction of the original, untokenized text by signaling where no space should be inserted.

Guidelines specific to the Romanian language and based on tokenization conventions used in previous Romanian linguistic resources developed in the UD framework were designed. As mentioned, we decided that the only compounds treated as multi-token words and to be disambiguated at token level are the ad-hoc, context specific compounds, which are not established dictionary entries (and therefore do not have an individual WSD label), but are formed productively as needed and typically exhibit compositional meaning: e.g., *sud-vest* "south-west": a directional compound formed as needed; *sud-dunărean* "south-Danubian": created to specify a geographic relation; *ruso-turc* "Russo-Turkish": combines two national/ethnic adjectives for a specific context. In such cases, the ELEXIS guidelines recommend splitting the compound while preserving compound information by specifying its span (see tokens 19–21 in Table 1).

Other hyphenated sequences are treated case by case, as follows:

- Compounds recorded as words (not locutions) in Romanian dictionaries are treated as one single token: e.g. *prim-ministru* "prime minister", *nou-născut* "newly born", *fast-food*, *mass-media*, *an-lumină* "light-year", *auto-raportat* "self-reported", *așa-numitul* 'so called' "the so called", etc.;

3 https://github.com/ufal/udpipe
4 romanian-rrt-ud-2.12-230717
5 https://universaldependencies.org/treebanks/ro_rrt/index.html

19-21	sud-vest	_	_	_	_	_	_	_	_
19	sud	_	_	_	_	_	_	_	SpaceAfter=No
20	-	_	_	_	_	_	_	_	SpaceAfter=No
21	vest	_	_	_	_	_	_	_	_

Table 1
Multi-token word treatment in ELEXIS

- Names of places are also treated as one token: e.g. *Nagorno-Karabah*, *Cluj-Napoca*;
- Semi-contextual compounds (whose tokens have independent meaning but the compound acquires a specific meaning of its own in specific contexts) are split: e.g., *Dunning-Kruger* → Dunning, -, Kruger , *unu-la-mai (mulți)* 'one-to-more' → unu, -, la, -, mai, (, mulți,);
- Opaque uses of hyphens: whenever it is not clear what the role of the hyphen is, we treat the elements it links as different tokens: e.g., *F-35B Lightning II* → F-35B, Lightning, II;
- Hyphen as inflection marker: split the inflexion from the root: *live-ului* "to the live" → live, -ului;
- Intervals marked by hyphen are split: e.g., *1521–26* → 1521, –, 26.
- Foreign compound words are also split: *Ready-to-wear* → ready, -, to, -, wear.

In Romanian, the hyphen plays an important role in marking contractions, resulting from cliticization and vowel elision. In such structures, the hyphen signals the phonological dependency of reduced functional elements (e.g., pronouns, auxiliaries) on adjacent lexical hosts. All contractions are split in the tokenisation process.

In structures involving pronouns, the spliting attaches the hyphen to the pronouns, which is the one affected by the vowel ellision:

- Clitic pronouns attached to verbs: *dă-mi cartea* "give me the book" -> dă, -mi;
- Auxiliary + pronoun contractions: *l-am văzut* 'him-have seen' "(I) have seen him";
- Clitic combinations: *mi-l dai* 'me-him give' "you give it to me" ⟶ mi, -l, dai; *ți-o dau* 'you-her give' "I give it to you" ⟶ ți,-o, dau.

Vowel ellision contractions can involve other functional elements like prepositions (*printr-o cunoștință* 'through-a acquaintance' "through a known person"), negation (*n-a fost* 'not-has been' "It hasn't been"), conjunctions (*c-ai venit* 'that-have come' "that you have come"). In all cases of vowel elision, the hyphen stays with the word from which the respective vowel dropped.

There are contractions when the hyphen does not replace any missing sound, like the gerund+clitic constructions, in which case the hyphen stays with the word that is not found in that form outside the respective construction: *Adaptându-se* → adaptându-, se (*adaptându* cannot occur but with the hyphen, it is not an independent form; its corresponding independent form is *adaptând*.

There are contextual special cases with specific rules when the principle of consistency is applied:

- In *văzându-și* 'seeing-Refl.', none of the words can occur independently, but the hyphen is consistently attached to the reflexive pronoun *și* and we attach it similarly in such cases for consistency.
- In *de-a lungul* → de-, a, lungul: *de-* and *a* are both prepositions and both can occur independently, but the hyphen is consistently attached at the end of a preposition (as opposed to the front) and we attach it similarly for consistency.

The apostrophe can occur in the following situation:

- in consonant elision, when it remains attached to the word (*domnu' Ionescu* "mister Ionescu" the apostrophe marks the elision of "l").
- in foreign words or sequences: all the token in the sequence are split, except the ones attached by apostrophe: S*gt. Pepper's Lonely Hearts Club Band* → Sgt., Pepper's, Lonely, Hearts, Club, Band; *don't* → don't.

Other specific rules were designed for:

- Symbols and formulas:
  - All terms of mathematical formulas are split: e.g., 76% → 76, %; *-100* → -, 100; *(3+5)*5/8=5* → (, 3, +, 5,), *, 5, /, 8, =, 5.
  - chemical compounds and formulas are kept as one token: e.g. *-NO2* (keep the polarity +/- together with the formula), *HLA-DRB1*11*.
  - Currency symbols are split from values: e.g., *3$* → 3, $
- Units of measure: the units are split from the values but kept as one token: e.g., *100 km/h* → 100, km/h.
- Slashes are always treated as individual tokens: e.g., *NOS/NTR* → NOS, /, NTR; *și/sau* "and/or" → și, /, sau
- Domain specific terms are split: e.g. *Haemophilus influenzae tip B* → Haemophilus, influenzae, tip, B; *F-35B Lightning II* → F-35B, Lightning, II.
- NEs are split: e.g. *NGC 205* → NGC, 205; *Apollo 12* → Apollo, 12.
- Abbreviations are treated as one token: e.g., *FC* → FC; *d.Hr.* → d.Hr.
- IT terms (written without blanks) are kept as one token: e.g. *W32.MyDoom@mm*, *www.voluntim.go.ro*.

3.3 Lemmatization and morpho-syntactic annotation

Annotation with parts of speech, with their features and the syntactic (i.e. dependency) relations between them was also done automatically, with UDPipe (Straka 2018), trained on the model romanian-rrt-ud-2.12-230717 based on the Romanian Reference Treebank (Barbu Mititelu 2018). The resulting CONLL-U files were converted to TSV3 format compatible with INCEpTION (Klie et al. 2018) (version 23.1). Based on a user account, the morphologic annotation and lemmatization were manually checked and corrected.

A number of recurrent annotation errors were identified during the processing of the Romanian data, many of which were caused by homonymy across PoSes. Such ambiguities frequently affected automatic tagging and lemmatization, requiring subsequent manual correction. One common source of error involved the ambiguity between adjectives and participles, particularly in forms that are morphologically identical but functionally distinct depending on context. Examples include *ridicată* "lifted/raised" and *utilizate* "used", which may function either as adjectival modifiers or as past participles within verbal constructions. Without sufficient contextual disambiguation, automatic tools often assigned inconsistent or incorrect tags.

Another frequent issue involved homonymy between proper nouns and common nouns, especially when common nouns appear in geographical or astronomical contexts and require capitalization. For instance, *pământ* may refer either to "ground" or to the planet "Earth", *lună* may denote "month" or "Moon", and *damasc* may refer either to the textile "damask" or to the proper noun "Damascus" (a metonymy). Automatic annotation often fail to capture these distinctions, particularly when capitalization conventions or contextual cues are subtle.

Ambiguity between adverbs and nouns also generated annotation inconsistencies. Forms such as *seara* "in the evening"/"the evening" or *vara* "in the summer"/"summer" may function either as temporal adverbs or as nominal expressions. Automatic tagging frequently misclassified these items, especially in contexts where syntactic cues were limited.

Homonymy within the same part of speech but involving different grammatical properties also posed challenges. For example, Romanian verbs may display identical forms in both present and imperfect tenses, such as *bea* "he drinks" / "he was drinking" and *beau* "they drink" / "they were drinking". Similarly, certain clitic pronouns share identical forms across grammatical cases, including accusative and dative, as in *le*, *îi*, *i-*, and *-le*, which complicated automatic morphological disambiguation. Another source of ambiguity involved nouns with identical forms in singular oblique case and plural direct case, such as *companii* "companies" and *soluții* "solutions", leading to inconsistent morphological analysis.

Abbreviations represented an additional challenge, as most were initially identified as proper nouns by the automatic annotation tools and subsequently corrected manually. Examples include *IT*, *MTV*, *TBC*, and *EBBA*, which required contextual interpretation to assign appropriate part-of-speech and entity labels.

Lemmatisation also revealed several systematic difficulties. One recurrent issue concerned neuter nouns appearing in plural form, which were incorrectly lemmatised with a feminine singular ending *-ă*. Examples include *tarifă* instead of *tarif*, *crateră* instead of

crater, and *metală* instead of *metal*. Another frequent error involved plural forms ending in *-i*, which were mistakenly interpreted as plural forms in *-uri* and truncated during lemmatisation. For instance, *armuri* was lemmatised as *arm* instead of *armură*, and *naturi* as *nat* instead of *natură*. These errors highlight the limitations of automatic lemmatisation for Romanian, particularly in the presence of morphological ambiguity and irregular inflectional patterns, and underscore the importance of manual validation in ensuring annotation accuracy.

3.4 Semantic annotations: MWEs

A further annotation level of the corpus is the semantic one. At the moment of this writing the corpus was automatically annotated with MWEs. These (i) are word combinations made up of at least two lexicalized components, (ii) have a neutral form which represents a weakly connected graph, and (iii) display some orthographic, morphological, syntactic and/or semantic idiosyncrasy with respect to what is considered general grammar rules of the respective language[6].

The types of MWEs currently represented in the Romanian corpus follow the PARSEME guidelines 2.0 and are as follows:

- Verbal MWEs:
  - 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 MWEs:
  - 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 MWEs:

6 We use the notion of MWE as defined within PARSEME guidelines: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=010_Definitions_and_scope/020_Multiword_expressions, last accessed April 20, 2026.

  - 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 MWEs:
  - 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 annotation was carried out using a lexicon of MWEs extracted from the previously annotated and manually validated PARSEME-Ro corpus (Barbu Mititelu et al. 2025). A total of 7,268 unique surface-form occurrences of MWEs, together with their corresponding PARSEME labels, were automatically retrieved from the corpus and subsequently matched against ELEXIS-Ro. The matching procedure allowed for intervening tokens between the components of a MWE in order to account for discontinuous realizations in context. To balance the trade-off between permitting gaps (which generate a large number of candidates requiring subsequent manual filtering) and prohibiting them (which results in lower recall and increases the need for manual annotation), we opted to allow up to two intervening tokens – a compromise intended to capture most long-distance dependencies while keeping noise at a manageable level. A total of 3,018 MWEs were automatically identified in ELEXIS-Ro, and their distribution according to PARSEME labels is presented in Table 3.

4. Statistics on the ELEXIS-Ro corpus

The corpus has 2024 sentences containing 35,316 words and the average sentence length is 17.45 words.

The distribution of tokens according to their part of speech can be seen in Table 2. This show the dominance of nominal structures (see the number of nouns, proper nouns, adjectives and pronouns in the corpus). Sentences are not too complex, as the average number of verbs per sentence is 1.5, while the number of subordinate conjunctions is also low (234).

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

Table 2
POS distribution in ELEXIS-Ro. The meaning of the abbreviation is as follows: ADP = prepositions, PUNCT = punctuation, ADJ = adjective, AUX = auxiliary verb, DET = determiner, PROPN = proper noun, ADV = adverb, PRON = pronoun, NUM = numeral, CCONJ = coordinating conjunction, PART = particle, SCONJ = subordinating conjunction, X = other, a residual class, INTJ = interjection, SYM = symbol

MWE type	Number	MWE type	Number
VID	79	NV.LVC.cause	4
LVC.full	18	AdjID	347
LVC.cause	2	AdvID	845
IRV	177	AV.VID	6
IAV	252	AV.IAV	30
NID	93	DetID	17
PronID	73	AdpID	863
NV.VID	1	ConjID	206
NV.LVC.full	1	IntjID	4
	TOTAL	**3018**	

Table 3
Number of MWEs by label in ELEXIS-Ro

5. Conclusions and future work

This paper has presented the extension of the ELEXIS multilingual corpus through the addition of Romanian as a new language component. The integration process was carried out primarily through automatic methods, followed by systematic manual validation to ensure annotation quality and cross-linguistic consistency.

At the current stage, manual revision has been completed for the translation, tokenisation, and morphological annotation layers. Observations regarding the types of errors encountered, the challenges specific to Romanian, and the strategies adopted for correction have been discussed in detail. These findings contribute to improving both the Romanian

component and the overall multilingual consistency of the ELEXIS corpus, while also providing insights into the interaction between automatic processing and manual validation in multilingual corpus development.

A key direction for future work concerns the manual validation and refinement of the automatically annotated MWEs in the Romanian component of the ELEXIS corpus. Although the current annotation follows the PARSEME 2.0 guidelines and relies on automatic identification procedures, previous studies have shown that such approaches inevitably introduce both false positives and false negatives, due to the structural variability and context-dependent interpretation of MWEs (Mititelu et al. 2026). In line with established methodologies, we plan to perform 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 types. Particular attention will be paid to ambiguous cases, homonymous expressions, and constructions whose MWE status depends on contextual interpretation.

The validation process will involve multiple trained annotators and will follow a cross-validation protocol inspired by previous annotation campaigns. Each candidate MWE will be independently assessed by at least two annotators, applying the PARSEME decision tests and annotation guidelines. Inter-annotator agreement will be measured in order to evaluate annotation consistency, and disagreement cases will be further analysed through adjudication procedures and expert discussion until consensus is reached. In addition, the validation stage will include manual correction of the automatic annotation at corpus level, such as removing incorrect annotations, adjusting MWE boundaries, and identifying missing expressions not captured automatically. This step is expected to substantially improve the quality and reliability of the Romanian dataset and to provide a solid basis for subsequent layers of annotation and cross-linguistic comparison within the ELEXIS framework.

The syntactic annotation of the corpus will also undergo manual validation by trained linguists familiar with the UD annotation principles and dependency relations inventory.

Given the importance of the semantic aspect in the design and development of the ELEXIS corpus, annotation of words and MWEs with senses will also be done. We envisage transfer of senses via aligned wordnets from languages that use wordnets as sense repositories for their sense annotation. This will naturally also be followed by manual validation.

The Romanian component of the ELEXIS corpus is part of the release of the new ELEXIS version 2.0 and is available with a permissive license.[7]

Acknowledgements

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 COST (European Cooperation in Science and Technology).

7 The corpus is available at https://vlo.clarin.eu/record/https_58__47__47_hdl.handle.net_47_11356_47_2101_64_format_61_cmdi?0.

References

Baisa, Vít, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, pages 2799–2803, European Language Resources Association (ELRA).

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

Barbu Mititelu, Verginica, 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 Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 113–125.

Barbu Mititelu, Verginica and Elena Irimia. 2016. Linguistic data retrievable from a treebank. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016), Sofia, Bulgaria*, pages 19–27, Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Barbu Mititelu, Verginica, Dan Tufiş, and Elena Irimia. 2018. The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*, European Language Resources Association (ELRA).

Ceauşu, Alexandru. 2008. Colectarea şi procesarea documentelor româneşti ale corpusului jrc-acquis (collecting and processing the Romanian documents of the JRC-Acquis corpus). In *Lucrările atelierului Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române (Proceedings of the Workshop Linguistic Resources and Tools for Processing the Romanian Language), Iaşi, Romania*.

Diewald, Nils, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP architecture — diving in the deep sea of corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, pages 3586–3591, European Language Resources Association (ELRA).

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

Ion, Radu, Elena Irimia, Dan Ştefănescu, and Dan Tufiş. 2012. ROMBAC: The Romanian balanced annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey*, pages 339–344, European Language Resources Association (ELRA).

Irimia, Elena, Verginica Barbu Mititelu, Radu Ion, Vasile Păiş, Maria Mitrofan, and Dan Tufiş. 2026. CoRoLa version 2.0: Corpus Enrichment and a New Annotation Level. In *Proceedings of the 12th Workshop on the Challenges in the Management of Large Corpora*, Language Resources Association (ELRA).

Klie, Jan-Christoph, 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 (COLING 2018)*, pages 5–9, Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X, Phuket, Thailand: Papers*, pages 79–86.

Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. Using textual and lexical resources in developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, 7:147–161.

Krstev, Cvetana, 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, Torino, Italy*, pages 106–114, ELRA and ICCL.

Macoveiciuc, Monica and Adam Kilgarriff. 2010. The RoWaC Corpus and Romanian WordSketches. In *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Romanian

Academy Publishing House, Bucharest, Romania, pages 151–168.

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

Martelli, Federico, 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. In *Proceedings of Electronic Lexicography in the 21st Century Conference*, pages 377–395, Lexical Computing CZ s.r.o.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Mititelu, Verginica, Mihaela Cristescu, Elena Irimia, and Carmen Mîrzea Vasile. 2026. Two birds with one stone: Annotating Romanian multiword expressions with an eye to the PARSEME 2.0 guidelines applicability. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), Rabat, Marocco*, pages 66–74, Association for Computational Linguistics.

Pedersen, Bolette, 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), Tórshavn, the Faroe Islands*, pages 11–18, Association for Computational Linguistics.

Savary, Agata, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, 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, Torino, Italy*, pages 372–382, ELRA and ICCL.

Straka, Milan. 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, Brussels, Belgium*, pages 197–207, Association for Computational Linguistics.

Straka, Milan, 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.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey*, pages 2214–2218, European Language Resources Association (ELRA).

Tufiş, Dan and Elena Irimia. 2006. RoCo-news: A hand validated journalistic corpus of Romanian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*, European Language Resources Association (ELRA).