CoRoLa version 2.0: Corpus Enrichment and a New Annotation Level

Elena Irimia, Verginica Barbu Mititelu, Radu Ion, Vasile Păiș, Maria Mitrofan, Dan Tufiș

Research Institute for Artificial Intelligence, Romanian Academy, Romania

{elena,vergi,radu,vasile,maria,tufis}@racai.ro

Abstract

The paper gives an overview of the recent developments in the enrichment of the reference Corpus of Contemporary Romanian (CoRoLa), within on-going international projects. Statistics of the newly acquired data, work methodology and work towards inclusion of a new annotation layer, the syntactic one, are detailed. We briefly present RODNA, an updated Romanian text processor with state-of-the-art performance on POS tagging, lemmatization and dependency parsing that will be used to populate the syntactic layer of CoRoLa.

Keywords: Romanian, corpus, NLP pipeline, RODNA

1. Introduction

Language resources in the form of large language corpora are still valuable assets provided that they are carefully curated and offer access to texts, metadata and annotation information inaccessible to the Large Language Models (LLMs) already widely available. In the current context, when Artificial Intelligence generated texts have become ubiquitous, corpora that guarantee the originality of their content and offer access to their content have high chances to become valuable repositories of the natural languages characteristics.

We present below the Reference Corpus of Contemporary Romanian (CoRoLa) (Barbu Mititelu et al., 2018), which was launched in 2017 and now is being under further quantitative and qualitative enrichment: new texts are collected, the design of the metadata is adjusted to reflect the newly added types of texts, and a new annotation level is added to the whole corpus. This will represent version 2.0 of CoRoLa.

Section 2 below describes the main characteristics of CoRoLa version 1.0, at the moment of its release in 2017, while the steps taken towards its version 2.0 are described in Section 3. This details the contexts in which new texts are collected and the methodology adopted for this mainly automatic collection. The major step taken in the corpus development is its syntactic parsing, using the dependency grammar, and this is the focus of Section 4, before concluding the paper.

2. CoRoLa 1.0

The development of a reference corpus to reflect the contemporary Romanian language was a priority project of the Romanian Academy, carried out by two of its institutes (Research Institute for Artificial Intelligence from Bucharest and the Institute for Computer Science from Iași). However, this was a national wide endeavour, as, on the one hand, collecting texts meant (and still means nowadays, given the lack of adaptation of the legislation to the evolution of technology and research) contacting publishers, media representatives and other decision makers in order to get access to the data. On the other hand, for their processing, human resources were needed, thus universities around the country were also contacted and they agreed to have their students involved in metadata creation and data cleaning.

Besides national bodies, the project also had an international component: it was due to the DRuKoLA project[1] (funded by the Alexander von Humboldt Foundation) that CoRoLa benefited from a reliable infrastructure to index its content and offer query-based access to it, i.e. the KorAP Corpus Query Platform (Diewald et al., 2016).

2.1. Design of CoRoLa

At the moment of its development, the corpus was meant to answer needs of several communities: linguists, for which the corpus is a source of various language phenomena attestation, offering a glimpse of their frequency, of the characteristics of various language styles, domains, etc.; language engineers, for which the corpus was a source of word embeddings (see below for those extracted from CoRoLa); the public, who can find here original, natural uses of various words, alongside their collocations.

As a work methodology, we gathered texts from various sources (from books to online newspapers, from textbooks to poetry, etc.) in various formats (PDF, DOC(X), MP3, WAV files), and both automatically and manually. Most of the collected data could be processed, cleaned and added to the corpus, while a small part of it proved unusable, as text could not be extracted from some files.

[1] https://www.ids-mannheim.de/digspra/pb-s1/projekte/drukola/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 91–97. Palma, Mallorca. May 11, 2026. ©ELRA Language Resources Association, 2026

2.2. Statistics of CoRoLa v. 1.0

The first version of CoRoLa comprised, in its written component, 1,257,752,812 tokens, unevenly distributed across multiple language styles (legal, administrative, scientific, journalistic, imaginative, memoirs and blog posts), four domains (arts and culture, nature, society, science) and 71 subdomains. The corpus was processed with the in-house tool TTL (Ion, 2007) for sentence splitting, tokenization, morpho-syntactic annotation and lemmatization, achieving an accuracy of approximately 97.5% (Tufiș et al., 2008). Documents were indexed in a local instance of the KorAP[2] corpus query and analysis platform (Diewald et al, 2019). Word embeddings[3] (Păiș and Tufiș, 2018) and frequency lists[4] computed on CoRoLa were also made public.

Moreover, the corpus has an oral component containing about 300 hours of recordings with transcriptions, which is also available for querying[5].

If at the moment of its launching, in 2017, a corpus of 1.2 billion tokens, entirely I(ntellectual)P(roperty)R(ights)-cleared, manually classified and validated, was considered a remarkable achievement, today this size is no longer impressive. At the same time, compliance requirements with IPR regulations remain stringent, and therefore the difficulties in collecting relevant data are still considerable. However, efforts to obtain usage rights for new data have proven successful, as shown below.

In what follows, we present new high-quality datasets that have been included or will be included in version 2.0 of CoRoLa. The data is both extensive and highly diverse, improving the initial distribution across linguistic styles and domains. In addition, new corpus processing tools, together with the inclusion of additional language varieties, enable broader investigations and extended applications.

2.3. Uses of CoRoLa

Over time, CoRoLa has supported:

- linguistic research (offering quality empirical data for theoretical studies: Ștefănescu, 2019; Ștefănescu et al., 2020; Giurgea, 2024; Bîlbîie, 2025; Vasileanu and Niculescu-Gorpin; 2025),
- the development of comparable corpora within projects such as: DruKoLa, which devised a pilot German-Romanian comparable corpus for contrastive linguistic analysis (Kupietz et al., 2019), CURLICAT[6] (Váradi et al., 2022), in which 3,042 documents from CoRoLa were IPR-cleared for further distribution by getting back to text providers with new agreement proposals
- Natural Language Processing tasks, such as Named Entity Recognition: CoRoLa-based embeddings were used to enhance a general named entity recognizer for Romanian implemented using Conditional Random Fields (CRF), based on the Stanford NER (Finkel et al., 2005) software package.

3. First Steps into CoRoLa 2.0

Recent international projects offered the perfect opportunity to resume corpus expansion, with particular care to exclude AI-generated content by restricting web crawling to texts published earlier than 2023.

3.1. MARCELL

In the project *Multilingual Resources for CEF.AT in the legal domain*[7] (MARCELL), funded by the Connecting Europe Facility, a large quantity of Romanian legal texts (163,000 files, 4,434,000 sentences, 412,000,000 tokens, which represent the body of national legislation ranging from 1881 to 2021) was collected and, given that such texts are not under any usage restrictions, they can be easily added to CoRoLa. All the texts were obtained via crawling from the public Romanian legislative portal[8]. We have not distinguished between "in force" and "out of force" laws because it is difficult to do this automatically and there is no external resource to use to distinguish between them. The texts were extracted from the original HTML format and converted into TXT files, metadata was automatically created and the texts underwent automatic processing and morpho-syntactic annotation.

3.2. LLMs4EU

Large Language Models for the European Union[9] (LLMs4EU) is a European Commission-funded project bringing together 66 European partners, including companies and research institutions, under the coordination of ALT-EDIC[10]. The project seeks to ensure the availability of LLMs and the necessary tools for their deployment across all EU languages by building on existing European programs and expertise and promoting open-data access. Within this framework, our team contributes by collecting new corpus data and providing fine-tuning and evaluation datasets, including resources derived from CoRoLa.

A set of 30,584 legal documents were extracted from the same Romanian legislative portal. The collected texts are from the period 2022-2025

[2] https://korap.racai.ro/
[3] https://corolaws.racai.ro/word_embeddings/
[4] https://zenodo.org/records/7091535
[5] http://89.38.230.23/corola_sound_search/index.php
[6] https://curlicat-project.eu/
[7] https://marcell-project.eu/
[8] https://www.just.ro/
[9] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/43152860/101198470
[10] https://language-data-space.ec.europa.eu/related-initiatives/alt-edic_en

(therefore not overlapping the data from MARCELL) and contain 2,383,809 sentences comprising 77,572,697 tokens. These IPR-cleared texts were annotated using UDPipe (Straka et al, 2016), and indexed in CoRoLa under a dedicated LLMs4EU sub-corpus label.

Also as part of the LLMs4EU project, we have extracted a corpus of Romanian language doctoral theses from the national Integrated Educational Registry platform (REI)[11], spanning thirteen broad scientific domains. The corpus comprises a total of 12,523 doctoral theses and over 770 million words (see Table 1), reflecting a cross-section of Romanian academic production. The largest contributions come from medicine and interdisciplinary fields, followed by engineering and humanities (see Table 1 for the exact numbers of documents and words as per each domain represented in the data). Smaller yet significant contributions are represented by arts, economics, theology, natural sciences, social sciences, life sciences, security and defense, law, and computer science. The total number of collected documents and that of words are also presented in Table 1.

At present, these theses are published under the Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International license (CC BY-NC-ND 4.0), which permits redistribution with attribution but prohibits commercial use and the creation or distribution of derivative works. Since NLP (Natural Language Processing) operations, including tokenization, fine-tuning, and the construction of training datasets for large language models, constitute derivative transformations under the terms of this license, we are currently in the process of contacting the host institution in order to obtain explicit permission to apply such derivative operations on these materials and include them in CoRoLa.

Domain	# docs	# words
medicine	3,654	176,455,463
interdisciplinary	2,507	174,166,039
engineering	1,724	83,558,112
humanities	664	66,858,211
arts	752	50,330,127
economics	683	47,056,316
theology	338	38,794,193
natural_sciences	615	30,854,763
social_sciences	364	24,190,568
life_sciences	402	21,070,967
security_defense	267	19,997,678
law	166	19,453,414
computer_science	387	17,763,040
TOTAL	**12,523**	**770,548,891**

Table 1: PhD thesis statistics.

Within the same project, we have identified new sources of quality data and have signed protocols for data collection, storing and processing. One such protocol was signed with "G. Călinescu" Institute for Literary History and Theory[12] and 51 IPR cleared files (containing 17, 525,736 words) have been received that are currently preprocessed in order to be included in CoRoLa. The documents have academic content from the philology and literary theory field.

Another batch of IPR-cleared files come from publishing houses and contain 2,981,620 words.

Altogether, within LLMs4EU texts comprising 868,628,944 words have been collected so far. They are in different processing steps, but they are all on their way into CoRoLa.

3.3. ADAMo

In the bilateral Romanian-Moldovan project *Automatic Detection of AI-Generated Texts from Moldova and Romania*[13] (ADAMo) our aim is to develop a classifier capable of identifying Artificial Intelligence (AI)-generated texts with characteristics from any of the two varieties of the Romanian language. Even though both countries have the same national language, Romanian, there are important differences between the variants spoken therein. For the first time, these specific features will be automatically identified at different language levels within this project, including the syntactic one, which is an annotation layer that will be added to CoRoLa within the ADAMo project.

At the moment, 12 million tokens (from the targeted 15 million) of high-quality, IPR-cleared (written, as well as oral) texts representative of the Moldovan variety of Romanian have already been collected in ADAMo and they will be added to CoRoLa, with adequate metadata, consistent with the ones for the texts already in the corpus.

Both the newly collected corpus and comparative texts extracted from CoRoLa will be used for developing the classifiers able to distinguish, on the one hand, between texts belonging to the two different language varieties, and, on the other, between original and AI-generated texts.

3.4. DeepNewDef

In the context of the project *Defending against deep fake news with large language and image models*[14] (DeepNewDef), we are building tools for

[11] https://rei.gov.ro/
[12] https://www.inst-calinescu.ro/
[13] https://www.racai.ro/p/adamo/index.html
[14] https://www.racai.ro/p/deepnewsdef/index.html

detecting fake news content (both text and images), considering the specifics from Romania and the Republic of Moldova.

In this context, the CoRoLa corpus will be used as a source of human-written text for training fake text detection models. Furthermore, as new content will be gathered throughout the project, original and IPR-cleared parts of it will be indexed and made available in CoRoLa.

4. Processing CoRoLa 2.0

The corpus will be reprocessed with RODNA[15] (**RO**manian **D**eep **N**eural networks **A**rchitectures, Ion, 2022), an actively developed, Romanian-specific NLP pipeline that performs sentence splitting, Romanian-aware tokenization, fine-grained Part-of-Speech (POS) tagging, lemmatization, and dependency parsing with Universal Dependencies (de Marneffe et al., 2021) dependency relations.

RODNA implements a Romanian-aware tokenizer that uses rules to correctly and consistently split dash-affixed clitics (e.g. "să-ți", "purtându-mi-l"), to recognize different types of numbers (real, percentages, integers grouped by three digits), time and dates (e.g. "12:15", "25/12/2025"), and abbreviations (e.g. "F.I.F.A.").

RODNA's lemmatizer uses a large Romanian lexicon[16] (more than 1.1M wordforms) in which each wordform is listed with its fine-grained POS tag (called a Morpho-Syntactic Descriptor or MSD[17]) and its lemma for that POS tag. Lemmatization is performed after POS tagging, so that we can get a set of lemmas for the pair wordform and MSD. If this set has more than one element, the most frequent lemma is used. When the word is not in the lexicon, it is assumed to be a content word (i.e. noun, verb, adjective or adverb) and lemmatization is obtained via the Romanian Paradigmatic Morphology (Irimia, 2009), which RODNA also implements.

For sentence splitting and POS tagging, RODNA uses a Romanian BERT model (Dumitrescu et al., 2020) to provide input embeddings for specialized neural networks heads that perform classification:

- Sentence splitting is done with a bidirectional LSTM (Long Short-Term Memory) network, that classifies a token as bearing the "end of sentence" mark or not.
- POS tagging is performed using the "tiered tagging" methodology (Tufiș and Dragomirescu, 2004): each token is first classified using a coarse-grained POS tagset, and then, the MSD is extracted from the Romanian lexicon, with a deterministic mapping from the pair (wordform, coarse-grained tag) to the MSD. If the word is not in the lexicon, the most probable MSD is assigned with a neural network that learns to map character embeddings of lexicon words to their possible MSDs.

RODNA's dependency parsing is realized in two steps:

1. construct the unlabeled dependency tree of the input sentence and
2. label each root-to-leaf path in this tree with Romanian UD dependency labels.

Step 1 learns a probability distribution of possible heads of the current token, as relative offsets (in number of tokens) to the left/right of the current token. Each token input is the BERT embedding for it, and we stack a bidirectional LSTM on top to learn a probability distribution of possible heads over a window of $\pm k$ tokens around the target token. Finally, using the Chu-Liu-Edmonds' algorithm for finding the maximum spanning tree in a directed graph (Chu and Liu, 1965; Edmonds, 1967) we obtain the unlabeled dependency tree of the input sentence.

Step 2 takes each root-to-leaf path in the unlabelled dependency tree and considers it an ordered sequence of BERT embeddings corresponding to tokens in the nodes. It learns to label each root-to-child edge by employing a unidirectional GRU neural network taking as input BERT embeddings and outputting a probability distribution over dependency labels.

RODNA has been trained on the "train" split of the Romanian Reference Treebank (RRT, Barbu Mititelu et al., 2016) and evaluated on the "test" split of this corpus, because the "dev" split was used to determine the best model to save, depending on the performance measure of the task on this split.

We compare RODNA to state-of-the art, multilingual text processing tools such as Stanza (Peng et al., 2020) and Trankit (Nguyen et al., 2021). Both text processors have been trained on the current version of the RRT "train" split and both of them use the "dev" split to choose and save their best models during training iterations.

Stanza also allows the integration of BERT embeddings; accordingly, it was trained using the same Romanian BERT model employed for RODNA. In contrast, Trankit relies exclusively on flavours of the XLM-RoBERTa model and does not support training with alternative BERT architectures.

Comparison is done by running all text processors on the raw text of the "test" split and letting each processor do sentence splitting, tokenization, POS tagging, lemmatization and dependency parsing. In order to compare them fairly, we align their outputs at sentence and token level with the gold standard, and compute accuracy for POS tagging and lemmatization, and Unlabelled

[15] https://github.com/racai-ai/Rodna

[16] https://github.com/racai-ai/Rodna/blob/master/data/resources/tbl.wordform.ro

[17] https://nl.ijs.si/ME/V6/msd/html/msd-ro.html

Attachment Scores (UAS) and Labelled Attachment Scores (LAS) for dependency parsing. In order to back up the claim that one processor is better than other processor, we use the McNemar's paired test to verify the null hypothesis that $n_{10} \approx n_{01}$, that is, the number of times the first text processor is correct when the second is not is about the same as the number of times the second text processor is correct when the first is not. n_{11} gives us the number of times both text processors are correct while n_{00} is the number of times neither is correct.

	n11	n00	n10	n01	Rodna	Stanza	Null
CG	14352	168	102	88	98.26%	98.16%	No
FG	14274	214	122	100	97.86%	97.71%	No
Lm	14337	88	205	80	98.85%	98%	Yes
US	12944	683	536	547	91.63%	91.71%	No
LS	12115	1177	684	734	87%	87.34%	No

Table 2: RODNA vs. Stanza on the "test" split of RRT

Table 2 shows the coarse-grained POS tagging accuracy (CG, with UD UPOS tags), the fine-grained POS tagging accuracy (FG, with MSDs), lemmatization accuracy (Lm), the UAS percentage (US) and the LAS percentage (LS). Underlined values are higher, but the Null column shows if the null hypothesis can be rejected or not, that is, if RODNA is significantly better than Stanza or the other way around, at a p-value of 0.05. Currently Rodna significantly outscores Stanza only when doing lemmatization because it uses the Romanian Paradigmatic Morphology when lemmatizing unknown words.

	n11	n00	n10	n01	Rodna	Trankit	Null
CG	14884	155	100	109	98.26%	98.32%	No
FG	14765	209	161	113	97.88%	97.53%	Yes
Lm	13997	117	1078	56	98.86%	92.16%	Yes
US	13138	897	631	582	90.3%	89.97%	No
LS	12310	1403	771	764	85.78%	85.74%	No

Table 3: RODNA vs. Trankit on the "test" split of RRT

With respect to Trankit, RODNA is significantly better when doing fine-grained POS tagging and especially when doing lemmatization: Trankit lemmatizer is subpar, when also tested with pre-trained models that it automatically downloads.

RODNA outputs files in the CoNLL-U format[18] which will be indexed in KoRAP, in a specific RODNA foundry, following KorAP's multi-layer indexing model (Diewald et al., 2016). This ensures that the lemmatisation, POS tagging and dependency parsing information, although accessible on different annotation layers, remains structurally aligned and can be queried within the same positional index.

By the end of 2026, we foresee the release of CoRoLa 2.0, with the following approximate counts, including the Moldovan variety of Romanian:

Domain	Style	No. of tokens
Law	Law	491,000,000
Science	Academic	790,000,000
-	Imaginative/Memoirs	3,000,000
-	Journalistic	13,000,000
TOTAL		1,297,000,000

Table 4: Projected counts for the CoRoLa 2.0 release.

5.Conclusions

The reference corpus of contemporary Romanian, CoRoLa, is an actively developed resource. All activities carried out towards its enrichment consider data quality and the possibility to use the processed texts in further resources, applications and downstream tasks development. CoRoLa is gaining language varieties representation (both at the written and the oral level), new types of texts (a particular type of academic one, namely doctoral theses) and is also harnessed in applications development specific to current days.

6. Acknowledgements

This work received support from: (i) a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV, (ii) the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union, (iii) a grant of the Ministry of Research, Innovation and Digitalization - UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (iv) NATO Science for Peace and Security Programme under grant id. G8648 (v) a grant of the Ministry of Education and Research, CCCDI - UEFISCDI, project number PN-IV-P8-8.2-NATO-SPS-2025-0005, within PNCDI IV. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

7. Bibliographical References

Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., and Perez, C. A. (2016). The Romanian treebank annotated according to universal

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

dependencies. In *Proceedings of the tenth international conference on natural language processing (hrtal2016)*.

Barbu Mititelu, V. (2018). Modern syntactic analysis of Romanian. In O. Ichim, L. Botoșineanu, D. Butnaru, M.-R. Clim, O. Ichim, & V. Olariu (Eds.), *Clasic și modern în cercetarea filologică românească actuală*, pp. 67–78. Iași: Publishing House of "Alexandru Ioan Cuza" University.

Barbu Mititelu, V., Tufiș, D., and Irimia, E. (2018). The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association. https://aclanthology.org/L18-1189/.

Bîlbîie, G. (2025) 'Multiple wh-questions in Romanian: A corpus-based approach', *AND CORPORA*, p. 33. https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Konferences/2025/GGC-10-ba.pdf#page=34

Chu, Y.J. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, *14*, pp.1396-1400.

de Marneffe, M.C., Manning, C., Nivre, J. and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics* 47(2): 255–308.

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)*, pp. 3586-3591.

Diewald, N., Barbu Mititelu, V., and Kupietz, M. (2019). The KorAP user interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique. 64*(3), 265-277

Dumitrescu, S., Avram, A. M., & Pyysalo, S. (2020). The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4324-4328).

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the national Bureau of Standards B*, *71*(4), pp.233-240.

Finkel, J.R., Grenager, T. and Manning, C.D., (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pp. 363-370.

Giurgea, I. (2024). Romanian double definites: The view from demonstratives. *Lingua*, *307*, p.103728.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD Thesis, Romanian Academy.

Ion, R. (Ed.) *Evaluating and User-Testing Rodna, A New Romanian Text Processing Pipeline*; Research report; Romanian Academy: Bucharest, Romania, 2022.

Irimia, E. (2009). ROG – A Paradigmatic Morphological Generator for Romanian. In *Vetulani, Z., Uszkoreit, H. (eds) Human Language Technology. Challenges of the Information Society. LTC 2007*. Lecture Notes in Computer Science, vol 5603. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04235-5_7

Kupietz, M., Cosma, R. and Witt, A. (2019). The Drukola Project. *Revue Roumaine de Linguistique. On design, creation and use of of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*, *64*(3), pp. 255-263.

Van Nguyen, M., Lai, V. D., Veyseh, A. P. B., & Nguyen, T. H. (2021). Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations,* pp. 80-90.

Păiș, V. and Tufiș, D. (2018). Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy, series A*, *19*(2), pp.403-409.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

Ştefănescu, A. (2019). The Use of Altminteri 'Otherwise'in Romanian: From Adverb to Textual Connector. In *Fuzzy Boundaries in Discourse Studies: Theoretical, Methodological, and Lexico-Grammatical Fuzziness* (pp. 287-313). Cham: Springer International Publishing.

Ștefănescu, A., Postolea, S. and Barbu Mititelu, V. (2020). The Romanian discourse markers *de altfel* and *de altminteri*: Patterns of use and core functions, *RRL*, 65(3), pp. 307–322.

Straka, M., Hajic, J., & Straková, J. (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)* (pp. 4290-4297). European Language Resource Association (ELRA).

Tufiș, D., Ion, R., Ceauşu, A., Ştefănescu D. (2008). RACAI's Linguistic Web Services. In *Nicoletta Calzolari et al. (Eds.) Proceedings of the 6th LREC, Marrakech, Morocco*, European Language Resources Association (ELRA).

Tufiș, D. & Dragomirescu, L. (2004, May). Tiered tagging revisited. In *Proceedings of the 4th LREC Conference. Lisbon, Portugal* (pp. 39-42). Language Resources Association (ELRA).

Váradi, T., Nyéki, B., Koeva, S., Tadić, M., Štefanec, V., Ogrodniczuk, M., Nitoń, B., Pęzik, P., Mititelu, V.B., Irimia, E. and Mitrofan, M. (2022). Introducing the CURLICAT corpora: seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 100-108). European Language Resources Association (ELRA).

Vasileanu, M. and Niculescu-Gorpin, A.-G. (2025) ‘Romanian libfixes in the making’, in Arndt-Lappe, S. and Filatkina, N. (eds.) *Dynamics at the lexicon-syntax interface: Creativity and routine in word-formation and multi-word expressions*. Berlin: De Gruyter, pp. 241–265.