

The Reference Corpus of the Contemporary Romanian Language (CoRoLa)

Verginica Barbu Mititelu, Dan Tufiş, Elena Irimia

Romanian Academy Research Institute for Artificial Intelligence
13 13 Septembrie Road, Bucharest 050711, Romania
{vergi, tufis, elena}@racai.ro

Abstract

We present here the largest publicly available corpus of Romanian. Its written component contains 1,257,752,812 tokens, distributed, in an unbalanced way, in several language styles (legal, administrative, scientific, journalistic, imaginative, memoirs, blogposts), in four domains (arts and culture, nature, society, science) and in 71 subdomains. The oral component consists of almost 152 hours of recordings, with associated transcribed texts. All files have CMDI metadata associated. The written texts are automatically sentence-split, tokenized, part-of-speech tagged, lemmatized; a part of them are also syntactically annotated. The oral files are aligned with their corresponding transcriptions at word-phoneme level. The transcriptions are also automatically part-of-speech tagged, lemmatized and syllabified. CoRoLa contains original, IPR-cleared texts and is representative for the contemporary phase of the language, covering mostly the last 20 years. Its written component can be queried using the KorAP corpus management platform, whereas the oral component can be queried via its written counterpart, followed by the possibility of listening to the results of the query, using an in-house tool.

Keywords: Romanian, reference corpus, annotation

1. Introduction

Language resources in the form of large corpora have been being created for more and more languages. We present here the results of a four-year project focused on the creation of a big corpus for contemporary Romanian language, called CoRoLa (corola.racai.ro). This has not been a singular effort: smaller previous or parallel projects (ANVSIB – <http://speed.pub.ro/anvsib>, SSPR – <http://dev.racai.ro/ti/wordpress/>, PARSEME – <https://typo.uni-konstanz.de/parseme/>) contributed to this project outcomes. In its turn, CoRoLa will contribute to other projects, larger ones, going beyond the national level (DruKoLA – <http://www1.ids-mannheim.de/direktion/kl/projekte/drukola.html>) and turning European. The CoRoLa corpus is now the reference corpus for contemporary Romanian. It is the largest one, containing 1,257,752,712 tokens for the written component and almost 152 hours of recordings for the oral component (the detailed structure is presented below). The texts cover all language styles, four major domains for which 71 subdomains were defined, thus ensuring a wide vocabulary coverage. The corpus can be reliably used as a basis for the creation of dictionary entries, grammar studies, other language reference materials, as well as for training and testing algorithms and systems for language processing.

CoRoLa has been jointly developed, as a priority project of the Romanian Academy, by two institutions: “Mihai Drăgănescu” Research Institute for Artificial Intelligence (from Bucharest) and the Institute of Computer Science (from Iaşi). Being located in different geographical and cultural regions of the country, the two partners could more easily contact texts providers from their vicinity, as unmediated, face-to-face contact and negotiations proved necessary for agreeing upon a protocol of collaboration after the correct understanding by our texts providers of the way texts are to be processed and further exploited when part of the corpus.

2. Related work

National corpora have been created for many languages: American English (Ide and Suderman, 2004), British English (<http://www.natcorp.ox.ac.uk/>), Bulgarian (Koeva et al., 2012), Croatian (Tadić, 2002), Czech (Křen et al., 2016), German (Kupietz and Lungen, 2014), Hungarian (Oravecz et al., 2014), Polish (Przepiórkowski et al., 2011), Russian (<http://www.ruscorpora.ru/en>), Turkish (Aksan et al., 2012), Eastern Armenian (<http://www.eanc.net/>), Welsh (Piao et al., 2016) and others.

Most of them are big (counting hundreds of thousands of words), with some being even huge (with over one billion words, see the Hungarian corpus, or even with tens of billions, as is the case of the German corpus). All of them reflect the language in the last twenty years, usually from various genres and domains; an exception is again the German corpus, which contains texts dating back to 1956 (Kupietz and Lungen, 2014).

Romanian corpora exist either as components of multilingual corpora or as monolingual resources. Within the former category we mention:

- the RO-JRC-Acquis (Ceaşu, 2008) – contains over 30 million words and reflects the law domain;
- Eur-Lex judgements corpus (Baisa et al., 2016) – contains over 17 million words and also reflects the law domain;
- EUROPARL corpus (Koehn, 2005) – contains almost 10 million words from the European Union Proceedings;
- OPUS (Tiedemann, 2012) – contains almost 300 million words and reflects mostly the legislative and administrative domains.

From the latter category, we mention:

- the RoCo-news corpus (Tufiş and Irimia, 2006) – contains around 7 million tokens and reflects the journalistic language;
- the RoWaC corpus (Macoveiciuc and Kilgarriff, 2010) – contains almost 45 million words gathered from the web;

- the Romanian balanced corpus ROMBAC (Ion et al., 2012) – contains about 36 million words, distributed rather evenly into five domains: journalistic, pharmaceutical and medical, law, literary history and fiction.

3. CoRoLa's Characteristics

The CoRoLa corpus stands out from these corpora due to its size, structure, origin and quality of texts. It is structured in files, each with associated metadata in CMDI format (most of them automatically created, but quite a lot manually), and annotated at several levels, as presented below.

The data collection and cleaning has been done as described by Tufiş et al. (2016). It involved both automatic and manual interventions on the data, for:

- bringing them into txt format (the most common formats in which we received them were pdf and doc(x)),
- separating texts (e.g., newspapers articles were separated in different files, just like chapters from edited books),
- recuperating text elements (paragraphs limits, removal of column marking newlines, words were recreated from their hyphenated form at the end of rows),
- removing unnecessary elements from the text (page numbers, headers, footers, footnotes, tables, figures, captions, etc.),
- diacritics insertion in the texts that lacked them or diacritics replacement (when non-standard ones occurred).

We present in this section several key characteristics of the texts included in CoRoLa: the time span they cover, originality, representativeness and I(ntellectual) P(roperty) R(ights)-clearance.

3.1 Time span

The corpus contains texts that could be collected in electronic format only, involving no OCR of scanned paper printed books. This condition limited the time coverage. The vast majority of texts reflect the language from the last twenty years. Texts from earlier time are legal ones and a few imaginative ones.

3.2 Originality

The vast majority of texts in CoRoLa are original ones: they are written in Romanian by native speakers. Only a few are translations. These belong to the legal style (translation of European legislation).

Another element of originality is the fact that the texts underwent no correction.

3.3 Representativeness

The corpus is representative for the contemporary language in that it contains texts from all language styles, major domains and many subdomains. The focus is only on the literary language. Nevertheless, the informal style can be found in the literary texts, although not explicitly marked in any way in the data and, thus, difficult to spot automatically.

3.4 IPR-clearance

During this project, great effort was invested in obtaining texts from their owners in a free of charge manner, given the scarce funding available. Major publishing houses, newspapers, magazines, news agencies, individual authors, and bloggers were contacted. Written protocols were signed with them so that we are allowed to freely obtain their (written or oral) texts, to store them on our servers, to process and annotate them and then to make them available for querying. We consider this a major achievement within this project, although this implies making (the largest part of) CoRoLa accessible only for querying and not for downloading.

The author's rights law in Romania does not have scope over legal and administrative texts. As such, they do not raise any storage or access problem.

4. Statistics

4.1 Written Texts

The distribution of the CoRoLa written texts across various styles is rendered in Table 1. We notice the unequal distribution of texts with respect to their style: legal texts are predominant.

Style	Number of tokens
legal	930,728,509
scientific	138,784,668
blogpost	53,704,460
journalistic	50,793,311
imaginative	47,727,438
administrative	17,759,778
memoirs	16,999,893
unclassified	1,254,755
TOTAL	1,257,752,812

Table 1: CoRoLa's texts distribution according to their style.

The distribution of texts according to the domain to which they belong is presented in Table 2.

Domain	Number of tokens
society	990,852,812
science nature	101,198,918
unclassified	93,511,926
arts and culture	70,510,600
nature	1,678,556
TOTAL	1,257,752,812

Table 2: CoRoLa's texts distribution according to their domain.

These domains are further classified into a various number of subdomains, as follows:

- architecture, art history, dance, design, fashion, film, folklore, literature, music, painting and drawing, poetry, sculpture and theater are the 13 subdomains of the arts and culture domain;

- environment, natural disasters, natural resources and universe are the 4 subdomains of the nature domain;
- administration, army, economy, education, entertainment, family, gossip, health, law, politics, religion, social events, social movements, sports, and tourism are the 15 subdomains of the society domain;
- archaeology, astronomy, biology, chemistry, constructions, criminalistics, engineering, ethnology, geography, geology, history, informatics, juridical sciences, linguistics, logics, mathematics, medicine, metrology, military science, oenology, pedagogy, pharmacology, philology, philosophy, physics, political sciences, psychology, religious studies and theology, sociology, standards, technics/technology are the 31 subdomains of the science domain.

We present below the data with the subdomains of each domain.

Subdomain	Number of tokens
literature	47,695,605
other	14,790,875
film	2,181,588
music	1,690,767
theatre	1,277,762
painting and drawing	1,153,013
architecture	776,447
folklore	445,277
art history	286,449
design	111,892
dance	42,407
fashion	24,545
sculpture	21,632
poetry	12,341
TOTAL Arts and culture	70,510,600

Table 3: CoRoLa's texts distribution according to their subdomain in the Arts and culture domain.

Subdomain	Number of tokens
other	1,003,512
environment	472,777
natural resources	110,728
universe	56,652
natural disasters	34,887
TOTAL	1,678,556

Table 4: CoRoLa's texts distribution according to their subdomain in the Nature domain.

Subdomain	Number of tokens
law	931,609,324
politics	19,820,835
other	11,977,554
economy	9,013,003
education	5,669,044
sports	3,538,550
religion	2,615,460
administration	2,193,158
gossip	1,676,574
health	862,677
entertainment	584,093
social events	519,484
tourism	436,682
family	319,614
social movements	14,205
army	2,555
TOTAL	990,852,812

Table 5: CoRoLa's texts distribution according to their subdomain in the Society domain.

Subdomain	Number of tokens
history	18,126,581
pharmacology	10,481,131
philology	10,269,512
medicine	10,093,872
sociology	8,228,897
geography	6,098,170
philosophy	4,868,934
psychology	4,727,630
political sciences	4,571,354
linguistics	4,116,164
religious studies and theology	3,734,212
other	3,196,889
pedagogy	1,906,633
biology	1,753,303
technics/technology	1,553,913
constructions	1,438,636
chemistry	1,302,917
juridical sciences	936,773
physics	741,234
mathematics	658,788
informatics	657,232
military science	457,075
archaeology	237,356

standards	195,617
astronomy	155,503
oenology	126,803
engineering	127,537
criminalistics	104,115
ethnology	98,828
geology	97,603
metrology	76,188
logics	59,518
TOTAL	101,198,918

Table 6: CoRoLa’s texts distribution according to their subdomain in the Science domain.

As a general remark, the corpus is unbalanced. We are still trying to enrich the less numerous categories, but, when collecting texts, it is hard to compete with a category without IPR restrictions (in this case, legal texts).

4.2 Oral Texts

The oral texts in CoRoLa are mainly professional recordings from various sources (radio stations, recording studios). They are accompanied by the written counterpart: the transcription either from their provider or made by us. As a consequence, different principles applied in their transcription.

Another part of the oral corpus is represented by read texts: read news in radio stations, texts read by professional speakers recorded in studios, and extracts from Romanian Wikipedia read by non-professionals, by volunteers, recorded in non-professional environments. In their case, the written component is provided by the sources, or was collected by us. An exception is the corpus RSS, compiled by Stan et al. (2011), which was enriched with ToBI-like annotation within our institute (Boroş et al., 2014).

The oral texts cover 151 hours 57 minutes and 21 seconds. However, not all of them have been processed yet. Their distribution according to the text type is given in Table 7 below.

Text type	Time (h:m:s)
News and radio interviews	119:52:33
News and fairy tales	03:44:00
Romanian Wikipedia	04:22:02
miscellanea	23:58:46
TOTAL	151:57:21

Table 7: CoRoLa’s oral text types and duration.

5. Annotation Levels

The annotation levels of the CoRoLa corpus are partially different for written and oral texts.

5.1 Written Texts

They are automatically sentence-split, tokenised and morpho-syntactically annotated, and lemmatised with an

in-house tool called TTL (Ion, 2007), having an accuracy of about 97.5% (Tufiş et al., 2008). We have not evaluated its accuracy on CoRoLa, though. The tagset of morpho-syntactic descriptions (MSDs) is compliant with MULTTEXT-EAST specifications (Erjavec, 2012).

A subcorpus of CoRoLa (9,522 sentences), containing samples of texts from the different styles, domains and subdomains existent in the corpus, is also consistently syntactically annotated and manually validated, following the UD principles (universaldependencies.org) and inventory of relations (Barbu Mititelu et al., 2016). However, we intend to annotate the rest of the CoRoLa corpus syntactically, as well, once we tune our recently developed parser (Ion et al., 2018) on corresponding styles, domains or subdomains.

Moreover, a part of the texts in the medical subdomain (18,000 sentences) is also annotated with medical terms of the type ANAT (anatomy or body parts), DISO (disorders), PROC (medical procedures) and CHEM (chemical substances) (Mitrofan, 2017; Mitrofan and Tufiş, 2018).

What is more, 51,500 sentences from the journalistic style were annotated with four types of verbal multiword expressions (namely, ID (idioms), LVC (light verb constructions), IRefIV (inherently reflexive verbs) and OTH (other)) for the PARSEME shared task on verbal multiword expressions identification (Savary et al., 2017).

5.2 Oral Texts

Up to now only a little more than half of the collected recordings (totaling about 300 hours) have been automatically preprocessed. Their written counterpart (counting 746,187 tokens) has been lemmatised, part-of-speech tagged and syllabified (Boroş and Dumitrescu, 2015); some allophones were identified and analyzed. The oral and the written texts are time aligned at the sentence, word and phoneme levels (Boroş et al., 2018) and this alignment was automatically encoded in separate files.

6. Access to CoRoLa

The corpus was publicly launched in December 2017. According to the protocols agreed upon and signed with texts providers, the corpus cannot be downloaded from our servers, but only queried. The results of queries are strings of a limited length, which makes the recovery of original texts impossible.

Access to written texts (<http://89.38.230.10:5555/>) is provided by the KorAP corpus query and analysis platform (Bański et al., 2014; Diewald et al, 2016), due to the running DruKoLA project (Cosma et al., 2016), in which Institut für Deutsche Sprache from Mannheim, where KorAP has been being developed, is our partner. KorAP allows for different type of linguistic searches (according to the annotation levels in the corpus, as well as combined levels), for creating virtual subcorpora (based on the metadata of the corpus). It allows regular expressions in query formulation. In Figure 1 we present a snapshot of the 37,464 results of the query `[orth=cel][orth=mai][drukola/m=pos:adverb]`, which searches for adverbs at the superlative degree.

A part of the corpus (about a fifth of it) is also available for search within the NLP-CQP web interface

(<http://89.38.230.23:1234/#>). This is meant to help those corpus users that cannot yet use a formal language for querying to formulate the query in controlled natural language (Romanian), then translates it in C(orpus)Q(uey)P(rocessor) (http://cwb.sourceforge.net/files/CQP_Tutorial/) format and retrieves the results. Figure 2 shows the query formulated in Romanian for finding 100 sentences in which the word “nu” occurs immediately after a verb. This query is translated into the CQP language: see the bottom of the figure.

At the moment, the oral texts can be queried (http://89.38.230.23/corola_sound_search/index.php) only via their transcription followed by rendering the relevant aligned speech segment. This is ensured by the in-house Oral Corpus Query Platform (OCQP) created by Vasile Păiș. It allows searching for lemmas or occurring words and morphological restrictions on them. The results are either sequences of 5 words or the whole sentences in which the specified form occurs. Besides the written form, the user also gets displayed the oral files, both for the searched for form and for the whole sentence to which it belongs. In Figure 3 we exemplify with the search of the word “copilărie” (En. “childhood”) and a few of the results obtained.

7. Sustainability

Although CoRoLa is already accessible to the users, its enrichment continues, targeting a balanced distribution of texts from various perspectives: text style, domain or subdomain, written and oral components. Two new projects (ReTeRom – Resources and technologies for developing human-machine interfaces in Romanian, and Heimdallr – Real-time Keyword Spotting in Telephone conversations) will enhance its oral component. In the perspective of creating EuReCo (Kupietz et al., 2017), CoRoLa can be enriched and exploited in harmony with other European corpora joining this initiative.

8. Conclusions and further work

CoRoLa is the result of a four-year effort, in which texts were collected from various sources, with the stated goal of going beyond what the web can offer. Its realization was possible due to the contribution of both individual people and juridical entities, all of them IPR owners. We are grateful to them all. Further data cleaning is necessary and has already started. Annotation quality needs to be assessed and a bootstrapping mechanism for its improvement will be applied. Moreover, further levels of annotation are targeted, and the syntactic one is a priority for the team.

Empirical studies of the Romanian language are now possible at various linguistic levels, in different styles, domains and subdomains. For a statistical analysis of the language we can also offer access to word embeddings (Păiș and Tufiş, 2018) and to n-grams of various sizes, as this does not constitute a breach of the protocols signed with texts providers.

9. References

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U., Yilmazer, H., Kurtoğlu, Ö., Atasoy, G., Öz, S., Yildiz, I. (2012). Construction of the Turkish National Corpus (TNC). In Proceedings of the Eight

- International Conference on Language Resources and Evaluation (LREC 2012). İstanbul. Türkiye.
- Baisa, V., Michelfeit, J., Medved', M., Jakubiček, M. (2016). European Union Language Resources in Sketch Engine. In *The Proceedings of tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA). Portorož, Slovenia.
- Bański, P., Diewald, N., Hanl, M., Kupietz, M., Witt, A. (2014). Access Control by Query Rewriting. The Case of KorAP. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA), pp. 3817-3822.
- Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., Perez, C.A. (2016) The Romanian Treebank Annotated According to Universal Dependencies. In *Proceedings of The Tenth International Conference on Natural Language Processing (HrTAL2016)*.
- Boroş, T., Stan, A., Watts, O., Dumitrescu, S.D. (2014). RSS-TOBI - a Prosodically Enhanced Romanian Speech Corpus. In Calzolari, Nicoletta et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: ELRA, 316-320.
- Boroş, T., Dumitrescu, Ş. (2015). Robust deep-learning models for text-to-speech synthesis support on embedded devices. In *Proceedings of the 7th International Conference on Management of computational and collective Intelligence in Digital Eco-System (MEDES'15)*, 98-102.
- Boroş, T., Dumitrescu, Ş., Pais, V. (2018). Tools and resources for Romanian text-to-speech and speech-to-text applications. In *Proceedings of The Second Workshop on Multi-Language Processing in a Globalising World and The First Workshop on Multilingualism at the intersection of Knowledge Bases and Machine Translation*.
- Ceaşu, Al. (2008). Colectarea și procesarea documentelor românești ale corpusului JRC-Acquis. In Diana Maria Trandabăţ, Dan Cristea, Dan Tufiş (eds.), *Lucrările atelierului Resurse Lingvistice și Instrumente pentru Prelucrarea Limbii Române*, Editura Universităţii „Al. I. Cuza”, Iaşi.
- Cosma, R., Cristea, D., Kupietz, M., Tufiş, D., Witt, A. (2016). DRuKoLA - Towards Contrastive German-Romanian Research based on Comparable Corpora. In Proceedings of the fourth CLMC, *LREC 2016, Portoroz*, Slovenia, 28 May, European Language Resources Association (ELRA).
- Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P. and A. Witt (2016). KorAP Architecture – Diving in the Deep Sea of Corpus Data. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz / Paris: ELRA: 3586-3591.
- Erjavec, T. (2012). MULTEXT-East: morpho-syntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, March 2012, Volume 46, Issue 1, pp. 131–142.
- Ide, N., Suderman, K. (2004). The American National Corpus First Release. In Maria Teresa Lino, Maria

- Francisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silva (eds.) *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, 1681-1684.
- Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD Thesis, Romanian Academy.
- Ion, R., Irimia, E., Ștefănescu, D., Tufiș, D. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In *Proceedings of LREC 2012*, 339-344.
- Ion, R., Irimia, E., Barbu Mititelu, V. (2018). Ensemble Romanian Dependency Parsing with Neural Networks. (this volume).
- Koeva, S., Stoyanova, I., Leseva, L., Dimitrova, T., Dekova, R., Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling*, 1 (1), 65-110.
- Koehn, Ph. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, vol 5.
- Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářiková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondříčka, P., Zasina, A. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2522–2528.
- Kupietz, M., Lungen, H. (2014). Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: ELRA, 2378-2385.
- Kupietz, M., Witt, A., Bański, P., Tufiș, D., Cristea, D., Váradi, T. (2017). EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In Bański, P., Kupietz, M., Lungen, H., Rayson, P., Biber, H., Breiteneder, E., Clematide, S., Mariani, J., Stevenson, M., Sick, T. (eds.), *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017* including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 15-19.
- Macoveiciuc, M., Kilgarriff, A., 2010, The RoWaC Corpus and Romanian WordSketches, in D. Tufiș, C. Forăscu (eds.), *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Romanian Academy Publishing House, Bucharest.
- Mitrofan, M. (2017). Bootstrapping a Romanian Corpus for Medical Named Entity Recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 501-509.
- Mitrofan, M., Tufiș, D. (2018). BioRo: The Biomedical Corpus for the Romanian Language. (this volume)
- Oravecz, C., Váradi, T., Sass, B. (2014). The Hungarian Gigaword Corpus. In Calzolari, Nicoletta et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland, 26-31 May, 1719-1723.
- Păiș, V., Tufiș, D. (2018). Computing Distributed Representations of Words using the CoRoLa Corpus. *Proceedings of the Romanian Academy, Series A*, vol 19. No.1/2018.
- Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez R-M., Knight, D., Křen, M., Löfberg, L., Nawab, R., Shafi, J., Teh, P-L. and Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. *Proceedings of the LREC (Language Resources Evaluation) 2016 Conference*, May 2016, Slovenia.
- Pipa, S., Boroș, T. (2016) A Recurrent Neural Networks Approach for Keyword Spotting Applied on Romanian Language. In *Proceedings of The 12th International Conference "Linguistic resources and tools for processing the romanian language*. Editura Universității "Alexandru Ioan Cuza" Iași, vol. 12, 111-119.
- Przepiórkowski, A., Bańko, M., Górski, R.G., Lewandowska-Tomaszczyk, B., Łaziński, M., Pęzik, P. (2011). National Corpus of Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland, November 25--27, 2011, 259–263.
- Savary, A., Ramisch, C., Cordeiro, S.R., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, Valencia, Spain, 4 April, 31-47.
- Stan, A., Yamagishi, J., King, S., & Aylett, M. (2011). The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*. 53(3), 442-450.
- Tadić, M. (2002). Building the Croatian National Corpus. In *Proceedings of LREC*, Las Palmas, Spain, 29-31 May, 441-446.
- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*
- Tufiș, D., Irimia, E. (2006). RoCo-News - A Hand Validated Journalistic Corpus of Romanian, In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 22–28.
- Tufiș, D., Ion, R., Ceaușu, A., Ștefănescu 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., Barbu Mititelu, V., Irimia, E., Dumitrescu, S.D., Boroș, T. (2016). *The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language*, in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 23-28 May, Portorož, Slovenia, p. 2516-2521.

Annexes.

The screenshot shows the KorAP search interface. At the top, the search query is `[orth=cel][orth=mai][drukola/m=pos:adverb]`. Below the search bar, it indicates "in all Corpora" and "with PoliQarp". The search results are displayed as a list of text snippets, each starting with a URL and followed by a snippet of text. The word "cel mai" is followed by an adjective in bold, such as "cel mai greu", "cel mai bine", "cel mai important", "cel mai probabil", "cel mai bine", "cel mai greu", "cel mai bine", "cel mai bine", "cel mai profund", and "cel mai bine".

Figure 1. A snapshot of the 37,464 results of the query `[orth=cel][orth=mai][drukola/m=pos:adverb]`, which searches for adverbs at the superlative degree in CoRoLa with KorAP.

The screenshot shows a Romanian search interface. At the top, the query is "100 de fraze în care cuvântul 'nu' apare imediat după un verb". Below the query, there is a button "Tradu în CQP!". The interface also shows a section "Dorești analiza traducerii?" with radio buttons for "Da" and "Nu". The main content area displays the query processed with TTL, showing the query in CQP format: `100/100/Mc de/de/Spsa fraze/frază/Ncfp-n în/in/Spsa care/care/Fw3--r cuvântul/cuvânt/Ncmsry "/" /DBLQ nu/nu/Qz "/" /DBLQ apare/apărea/Vmip3s imediat/imediat/Rgp după/după/Spsa un/un/Timsr verb/verb/Ncms-n`. Below this, there are sections for "Expresii CQP de tip 'context'", "Expresii CQP de tip 'termen'", and "Expresii CQP de tip 'predicat'", each with a "click pentru analiză" button. The "context" section shows C1: `cut 100` and C2: `set Context s`. The "termen" section shows T1: `[(word = "nu")]` and T2: `[(pos = "V.*")]`. The "predicat" section shows T1: `SEQUENCEREVERSE1` and T2 (= T3): `[(pos = "V.*")] [(word = "nu")]`.

Figure 2. The query formulated in Romanian for finding 100 sentences in which the word “nu” occurs immediately after a verb and its translation into CQP.

Cuvânt ▾ = copilărie (opțional:AND) CTAG ▾ =

Afișare: Cuvinte Lema MSD CTAG | Context: 5 Cuvinte ▾

Caută!

Rezultatele căutării pentru " copilărie " (11 rezultate)

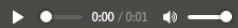
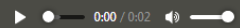
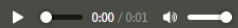
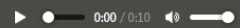
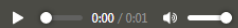
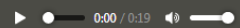
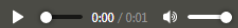
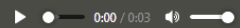
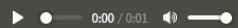
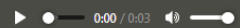
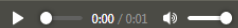
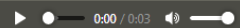
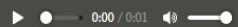
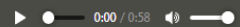
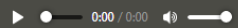
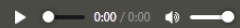
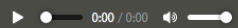
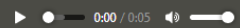
Context	Ascultare cuvânt	Ascultare frază
Din copilărie a înțeles ce	 0:00 / 0:01	 0:00 / 0:02
plăcute din copilărie .	 0:00 / 0:01	 0:00 / 0:10
încă de copilărie : fetița	 0:00 / 0:01	 0:00 / 0:19
avut o copilărie nefericită .	 0:00 / 0:01	 0:00 / 0:03
avut o copilărie nefericită .	 0:00 / 0:01	 0:00 / 0:03
avut o copilărie nefericită .	 0:00 / 0:01	 0:00 / 0:03
ceva din copilărie , și	 0:00 / 0:01	 0:00 / 0:58
știm din copilărie și care	 0:00 / 0:00	 0:00 / 0:00
mai fragedă copilărie , se	 0:00 / 0:00	 0:00 / 0:05

Figure 3. Some results of the search for the word “copilărie” (En. “childhood”) in the oral component of CoRoLa.