Chapter 31
Language Report Romanian

Vasile Păiş and Dan Tufiş

Abstract Since the previous META-NET report, there have been significant improvements (e. g., creation of a large Romanian national corpus, steady progress in written language technologies, LT, construction of a national LT portal for the Romanian language etc.), but things are far from what they should be. Support for LT and AI through national programmes is still modest, although there are signs of a more active involvement of policy makers in the strategic planning and funding programmes in this domain. Continued research is required to produce large language models, able to capture the characteristics of the Romanian language. Large language resources need to be created so that AI systems are able to learn from them.

1 The Romanian Language

The Romanian language which is an official language of the EU is also the official language of Romania. It is spoken by 19.4 million people in Romania and by about 3.5 million people in Moldova, where it is unofficially known as a Moldavian language. Speakers of Romanian in other European countries (Albania, Bulgaria, Croatia, Greece, Hungary, North Macedonia, Serbia, Ukraine and others) and communities of immigrants in Australia, Canada, Israel, Latin America, Turkey, USA and Asian countries total around 4 million Romanian native speakers.[1]

Romanian is an official language in the Autonomous Province of Vojvodina in Serbia. It is one of the languages spoken in the autonomous Mount Athos in Greece and a recognised minority language in Ukraine (Trandabăț et al. 2012). Romanian has four dialects: Daco–Romanian, Aromanian (about 500,000 speakers in Albania, Bulgaria, Greece and North Macedonia), Istro–Romanian (15,000 speakers in two small areas in the Istrian Peninsula, Croatia) and Megleno–Romanian (about 5,000 speakers in Greece and North Macedonia).

Vasile Păiş · Dan Tufiş
Romanian Academy, Romania, vasile@racai.ro, tufis@racai.ro

[1] https://en.wikipedia.org/wiki/Romanian_diaspora

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_31

The Romanian alphabet is based on the Latin script with five additional letters using diacritics (Ă, Â, Î, Ș, Ț and ă, â, î, ș, ț). Many digital texts are written without diacritics. The quotation marks use double low (left) and right marks („ and ”, respectively). However, especially in digital texts, the ASCII quotation mark character may be encountered. Dialogues are introduced using quotation dashes (–). The Oxford comma, used in certain English language documents, is considered incorrect in the Romanian language. In titles, only the first letter of the first word is capitalised, with the rest of the title making use of regular sentence capitalisation. Names of months and days, as well as adjectives derived from proper names are not capitalised, e. g., februarie (February), vineri (Friday), italian (Italian).

2 Technologies and Resources for Romanian

The availability of language-specific data has a direct impact on the quality of language-specific or cross-language tools. The availability of large pre-trained multilingual models that include representations for Romanian language, such as XLM-RoBERTa or mBERT, somewhat alleviates the problem of constructing compute-intensive contextual word representations. Nevertheless, monolingual representations such as RoBERT (Masala et al. 2020), DistilRoBERT (Avram et al. 2022), and ALR-BERT, led to increased performance of monolingual tools (Tufiș 2022). Static representations, such as CoRoLa-based word embeddings (Păiș and Tufiș 2018), are still used due to their lower compute requirements (Păiș and Tufiș 2022).

Word representations form only the basis of advanced language tools. In addition to language models, task-specific corpora are required to train and evaluate the tools. The vast majority of Romanian resources are multilingual, with some being bilingual, and only a few monolingual corpora exist. Compared to English, the available Romanian corpora represent around 10%. Available speech corpora with Romanian audio represent 5% of available English resources and about 50% when compared to neighbouring EU countries.

In spite of the reduced number of available language resources, applications for different NLP tasks exist for Romanian. These include lemmatisation, part-of-speech tagging, dependency parsing, named entity recognition, syllabification, speech recognition, text-to-speech, machine translation, punctuation restoration, terminology annotation, and text classification. The number of identified tools represents only 15% of the tools available for English.

Even if, in general, all LT fields are covered, certain fields are less developed or considered for the Romanian language by researchers and developers: language generation, dialogue management, multimodal corpus building, and social media aspects (including micro-blogging, social networks, and meme interpretation). Speech processing is much less mature than LT for written text, both in terms of corpora and instruments. Even though there has been much work on processing general Romanian language, more focus is needed for creating domain-specific resources and tools (especially for the biomedical, legal, economy and social media domains).

The Representative Corpus of Contemporary Romanian Language (CoRoLa)[2] (Tufiş et al. 2019) was created by the Romanian Academy as the largest IPR-cleared reference corpus of written and spoken Romanian. Texts cover four domains (arts and culture, science, society, nature), reflecting six styles (imaginative, journalistic, scientific, legal, administrative, memoirs) and different document types.

One of the largest Romanian speech corpora is RSC (Georgescu et al. 2020), containing 100 hours of audio files. The multilingual speech corpus VoxPopuli contains 83 hours of Romanian language speech. The speech component of the CoRoLa corpus (comprised of multiple smaller corpora together with additional audio files specifically obtained for inclusion in CoRoLa) totals 103 hours aligned with the text.

A number of Romanian LTs, covering different fields of research, are available within the RELATE[3] (Păiş et al. 2020) portal. The platform covers results derived from more than six national and international research projects.

3 Recommendations and Next Steps

Task-specific Romanian corpora (including multi-modal) are needed to enable new and complex language processing operations. In turn, these must lead to the development of new tools, finally working towards digital language equality. This requires dedicated long-term support at the national, regional and European levels. Furthermore, AI research should follow a human-centered approach. Biased or potentially harmful data in resources should be detected and addressed. This, together with following lawful and ethical principles, as well as robust implementations, should enable building Trustworthy AI (TAI)[4] applications for the Romanian language.

AI is an area of strategic importance and a key driver of economic development, providing solutions to many societal challenges. In this context, many EU countries prepared national plans for AI (e. g., the *Spanish National AI Strategy*[5] or the *French AI for Humanity*[6]). In Romania, however, there is currently no such national plan for AI. A strategy for AI[7] has been proposed recently within the RePatriot[8] project, but it was not adopted at national level. Furthermore, the strategy is not very concrete, it centres mostly on which Romanian sectors would benefit most from AI, and which steps are important for the process of developing Romanian AI initiatives, but it does not include any plans about how to accomplish these actions.

---

[2] https://corola.racai.ro

[3] https://relate.racai.ro

[4] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[5] https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/021220-ENIA.pdf

[6] https://www.aiforhumanity.fr/en/

[7] https://www.slideshare.net/MonicaIon1/strategy-romania-in-the-era-of-artificial-intelligence-rblrepatriot

[8] https://repatriot.ro

References

Avram, Andrei-Marius, Darius Catrina, Dumitru-Clementin Cercel, Mihai Dascalu, Traian Rebedea, Vasile Pais, and Dan Tufis (2022). "Distilling the Knowledge of Romanian BERTs Using Multiple Teachers". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 374–384. https://aclanthology.org/2022.lrec-1.39.

Georgescu, Alexandru-Lucian, Horia Cucu, Andi Buzo, and Corneliu Burileanu (2020). "RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 6606–6612. https://aclanthology.org/2020.lrec-1.814.

Masala, Mihai, Stefan Ruseti, and Mihai Dascalu (2020). "RoBERT – A Romanian BERT Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, pp. 6626–6637. DOI: 10.18653/v1/2020.coling-main.581. https://aclanthology.org/2020.coling-main.581.

Păiș, Vasile, Radu Ion, and Dan Tufiș (2020). "A Processing Platform Relating Data and Tools for Romanian Language". In: *Proceedings of the 1st International Workshop on Language Technology Platforms*. Marseille: European Language Resources Association, pp. 81–88. https://aclanthology.org/2020.iwltp-1.13.

Păiș, Vasile and Dan Tufiș (2018). "Computing distributed representations of words using the CoRoLa corpus". In: *Proceedings of the Romanian Academy Series A* 19.2, pp. 185–191.

Păiș, Vasile and Dan Tufiș (2022). *Deliverable D1.29 Report on the Romanian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-romanian.pdf.

Trandabăț, Diana, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiș (2012). *Limba română în era digitală – The Romanian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/romanian.

Tufiș, Dan (2022). "Romanian Language Technology – a view from an academic perspective". In: *International Journal of Computers Communications & Control* 17.1. DOI: 10.15837/ijccc.2022.1.4641.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei (2019). "Little Strokes Fell Great Oaks. Creating CoRoLa, The Reference Corpus of Contemporary Romanian". In: *Revue roumaine de linguistique* 64.3, pp. 227–240.