A Bird's-eye View of Language Processing Projects at the Romanian Academy

Dan Tufiș[1], Dan Cristea[2,3]
[1]Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy,
[2]Institute for Computer Science of the Romanian Academy
[3] "Alexandru Ioan Cuza" University of Iaşi
[1]13Calea 13 Septembrie, Bucharest, Romania
[2]2 T. Codrescu St, 700481, Iaşi, Romania
[3]11 Carol I Blv, 700506, Iaşi, Romania

[1]tufis@racai.ro, [2]dcristea@info.uaic.ro

Abstract

This article gives a general overview of five AI language-related projects that address contemporary Romanian language, in both textual and speech form, language related applications, as well as collections of old historic documents and medical archives. Namely, these projects deal with: the creation of a contemporary Romanian language text and speech corpus, resources and technologies for developing human-machine interfaces in spoken Romanian, digitization and transcription of old Romanian language documents drafted in Cyrillic into the modern Latin alphabet, digitization of the oldest archive of diabetes medical records and dialogue systems with personal robots and autonomous vehicles. The technologies involved for attaining the objectives range from image processing (intelligent character recognition for hand-writing and old Romanian documents) to natural language and speech processing techniques (corpus compiling and documentation, multi-level processing, transliteration of different old scripts into modern Romanian, command language processing, various levels of speech-text alignments, ASR, TTS, keyword spotting, etc.). Some of these projects are approaching the end, others have just started and others are about to start. All the reported projects are national ones, less documented than the international projects we are/were engaged in, and involve large teams of experts and master/PhD students from computer science, mathematics, linguistics, philology and library sciences.

Keywords: contemporary language reference corpus, speech corpus, natural language dialogue systems, intelligent character recognition, medical records archive

1. Introduction

Language technology is a discipline needed in more and more areas. The computational linguistics traditional domain has been extended towards dealing with vast and diversified quantities of linguistic data, both written and spoken. This extension requires putting together expertise from natural language processing (NLP), Internet-of-Things (IoT), robotics, image processing, cognitive computing, High Performance Computing (HPC), semantic web, social media and many others. In this article we will give a brief account on several projects at the Romanian Academy, their prospects and outline the challenges posed by each of them. The common characteristic of these projects, apart from being all conducted by teams affiliated to different institutes of the Romanian Academy, is the use of a solid background of language resources, tools and models, acquired along the last two decades and permanently enhanced. As such, we argue that a constant accumulation of human expertise, language resources and software utilities can bring about a level that allows for the development of extremely complex projects. Without this know-how, data and processing platforms, these enterprises would have not been feasible.

2. CoRoLa

The CoRoLa project is a priority program of the Romanian Academy and is developed by two institutes of the Romanian Academy: "Mihai Drăgănescu" Research Institute on Artificial Intelligence in Bucharest (RACAI) and the Institute of Computer Science of the Iaşi branch of the Romanian Academy (IIT), and voluntarily contributed by linguists from "Al. Philippide" Institute of Romanian Philology in Iaşi and many master and PhD students from "Alexandru Ioan Cuza" University of Iași, University "Politehnica" of Bucharest and the Bucharest University.

The project is concerned with the construction and maintenance of the reference corpus of the contemporary written and spoken Romanian as well as with the infrastructure to openly use it.

The corpus covers 5 large domains (arts & culture, society, science, nature and *others*[1]) which are further refined into 71 sub-domains - see (Barbu MItitelu *et al.*, 2018) for more details.

Currently, CoRoLa contains over 1,200,000,000 tokens of written texts and almost 152 hours of pre-processed speech recordings and may be queried at *corola.racai.ro* via three web interfaces (see Figure 1, 2 and 3 in the annex): KorAP (the entire text corpus is searchable), NLP-CQP (only a selection of about 35% from the entire text corpus is searchable) and OCQP (the oral corpus).

All the CoRoLa data is IPR cleared based on individual written agreements concluded with the data providers (IPR holders). The written texts are fully pre-processed (sentence split, tokenized, lemmatized and morpho-syntactically tagged). A small part of the corpus represents a core of a Reference Treebank dependency parsed (RoRefTrees) (approx. 10,000 sentences) and hand validated (Barbu-Mititelu *et al.*, 2016). In (Barbu-Mititelu *et al.*, 2016) there is a description of the previous RACAI

[1] "others" is a category for all documents that could not be definitely classified into the named categories.

dependency parser, based on MaltParser (Nivre *et al.*, 2007), trained on RoRefTrees. Recently, a new parser, SSPR (Semantics-driven Syntactic Parser (for Romanian)) has been developed. It is a feed-forward neural network ensemble parser developed for Romanian. The network architecture contains two hidden layers, the first with 100 neurons, and the second with 10 neurons. The activation function for both hidden layers is *tanh*. The parser is actually a classifier that combines the parsing decisions of three (the number is arbitrary) different parsers, namely MaltParser, RGB (Lei et al., 2014) and MATE (Bohnet, 2010) and using information from additional lexical, morpho-syntactic and semantic features builds the most likely dependency parse (see (Ion et al, 2018) for further details). SSPR performs better than any of the three individual parsers, with average scores of 83% LAS (labeled attachment score) and 88.44% UAS (unlabeled attachment score). Thus, SSPR is in the same accuracy class with the top 5 performers at the CONLL 2017 dependency parsing shared task for Romanian. This statement is backed up by the fact that the training, development and test data used for comparison were identical (RoRefTrees).

SSPR will be used to parse the entire content of the textual part of CoRoLa. The current textual content (which will continue to grow) is, quantitatively, well above the project initial promise (500 million tokens).

Recently, we constructed various sets of word embeddings from the CoRoLa corpus (Păiș and Tufiș, 2018), using different vector sizes (100, 200, 300, 400, 500 and 600) and frequency thresholds (1, 5, 10, 20, 50). The pre-trained CoRoLa word embeddings, analogy application and data sets may be downloaded from http://89.38.230.23/word_embeddings/. The user may play around with associations such as "însurat" (En. *married*, term used only for men) - "bărbat" (En. *man*) + 'femeie" (En. *woman*) and get "măritată" (En. *married*, term used only for women).

The situation is not equally good for the pre-processed spoken data, although the quantity of the collected raw speech data exceeds the promised 300 hours. The reason is that several sets of spoken material are not transcribed and some are not IPR-cleared yet (mainly children stories). The transcribed spoken data is pre-processed (sentence split, tokenized, lemmatized and morpho-syntactically tagged) and aligned with the speech data at sentence, word and phoneme levels. The interface allows a user to search for a word and hear its pronunciation or to listen to the entire sentence where it occurs.

The CoRoLa language data is accompanied by detailed meta-data information (CMDI compliant). The initial corpus management system was IMS Open Corpus Workbench, an open source medium (CWB, http://cwb.sourceforge.net/) but lately, based on a collaboration with IDS-Mannheim, we commuted on their KorAP (Banski et al., 2014) environment which appears to be rather insensitive (response time-wise) to the volume of the corpus. For the time being, KorAP does not offer facilities for querying the speech data, so we continue to use our own interface. However, as it will become available, all CoRoLa data (text and speech) will be searchable in a single unified environment (KorAP).

This collaboration, financed by the Humboldt foundation, joints forces from IDS Mannheim, University of Bucharest, IIT-Iași and RACAI-Bucharest in the DRuKoLA project (Cosma et al., 2016). There are further developments on this project, hopefully towards a pan-European initiative – EuReCo (Kupietz *et al.*, 2017*).*

The first phase of CoRoLa project (2014 – 2017) ended with an official public launching with presentations and demos, enjoying a high interest from the academic community, as well as from the media. The Romanian Academy decided that the CoRoLa project will be further continued and enhanced (with the same status of a priority program) within the next phase (2018-2020). The enhancement of the CoRoLa corpus will be pursued in parallel with the DRuKoLA project and, possibly, with EuReCo partnership.

3. Human-Machine Interfaces in Spoken Romanian

To compensate the shortage of the IPR-cleared speech data, we launched this year a national project involving the major local players in speech processing: University POLITEHNICA of Bucharest, Technical University of Cluj-Napoca, "A.I. Cuza" University of Iaşi and RACAI-Bucharest. All partners have developed small but clean corpora of transcribed pre-processed speech data (few hundred hours of speech recordings) and one objective of this project is to harmonize these speech data collections into a single larger common speech corpus that will contain at least 500 hours of fully processed speech recordings and which will be documented with the appropriate unified meta-data. The recordings represent, in general, read texts, but also multiple participants interviews and a few recorded hours or spontaneous speech. This new speech corpus will be included into CoRoLa. Other objectives are the development of an accurate, speaker independent, Automatic Speech Recognition (ASR) system (which we plan to use for transcribing the bulk of not yet transcribed records) and a high quality Text to Speech (TTS) system capable to generate expressive (emotional) speech, based on subjectivity marked-up text (Tufiș and Ștefănescu, 2012). This project aims at producing portable and interoperable much better ASR and TTS systems than the existing ones (Cucu et al. 2014), (Stan et all. 2013), (Boroș and Dumitrescu, 2015). The ASR and TTS systems will be integrated into a platform for trainable, generic and situated (script-based) dialogues.

This work is supported by a grant of the Ministry of Research and Innovation CCCDI-UEFISCDI, project cod PN-III-P1-1.2-PCCDI-2017-0818 within PNCDI III.

In parallel with this project we are working, together with experts from the Military Academy, on an experimental-demonstrative project (to be finished in 2018) for real time detection of keywords (Romanian language) in telephone conversations (*http://heimdall.racai.ro/*).

Keyword Spotting (KWS) is a technology that enables the detection of word occurrences within spoken language

from audio or audio/video streams. While it shares many aspects in terms of audio processing with speech recognition and transcription, it is fundamentally different from any restricted (grammar based) or language model (LM) based speech transcription system. The main difference between these technologies is that KWS does not require any language modeling and, provided with automatic grapheme-to-phoneme (G2P) capabilities, KWS is not in any way affected by out-of-vocabulary (OOV) words.

The acoustic models will be constructed based on the oral corpus in CoRoLa and an additional corpus of telephone recordings developed within this project. The data is annotated for speaker's gender (male-female), age interval (18-35/35-60) and regional accent (Bucovina, Moldova, Oltenia, Muntenia).

Though KWS is developed as a system for security applications, it is not restricted to this field. With minor adjustments the system can be used for call-center monitoring, business analytics, etc.

The experimental model will be a KWS software integrated with a VoIP laboratory network. The system will be able to detect keywords in live conversations in real-time. The proposed system starts from a proof of concept KWS developed by RACAI, and improves it on the following areas: acoustic models, enlarged corpora for training the acoustic models and real time performance.

This work is supported by a grant of the Ministry of Research and Innovation CCCDI-UEFISCDI, project code PN-III-P2-2.1-PED-2016-1within PNCDI III.

4. Speaking with Your Personal Robot

This is a complex project (ROBIN), user centered, aiming to develop software and services for interaction with assistive robots and autonomous vehicles. The project consortium, besides RACAI, includes experts in robotics and digital signal processing from the Polytechnics University of Bucharest and the Institute of Mathematics of the Romanian Academy, well known specialists in IoT technology from the University of Bucharest and Technical University of Cluj-Napoca and experts in designing autonomous vehicles from University "Dunărea de Jos" of Galați. ROBIN is an ensemble of five collaborating sub-projects, combining advanced techniques and technologies from AI, human-robot interaction, pervasive and Cloud computing, each of them having specific objectives. Language interaction is one of the challenges, as no robots can be talked to now in Romanian, neither any interaction, with appropriately equipped cars, may be conducted in Romanian.

The human-robot dialogues are situated ones (close-world based) and our involvement will be supported by previous results in the recently finished project ANVSIB (Boroș and Dumitrescu, 2017) on intelligent buildings equipped with IoT devices controlling the home appliances (TV set, heating system, drapes and interior blinds, main locks, etc.).

The situated dialogue component of ANVSIB project, enhanced with the new facilities to be developed in the previously mentioned generic dialogue platform will be adapted for this project. However, the major challenge is that the spoken commands will not be sent via a smart phone (as in ANVSIB project), but addressed directly to the robot, thus being affected by more surrounding noise.

The sub-project concerned with the development of intelligent software modules for hands-off driving, automated driving and a prototype of electric semi-autonomous vehicle benefits of a prototype DACIA car offered by an industrial partner (PRIME Motors Industry). Within this sub-project, our team will be responsible for the development of a module for voice interaction (Romanian language) between the driver and the car.

This work is supported by a 33 month grant of the Ministry of Research and Innovation PCCDI-UEFISCDI, project code PN-III-P1-1.2-PCCDI-2017-734 within PNCDI III.

5. Intelligent Character Recognition in Medical Records Processing

Automatic handwriting recognition is a major preoccupation for commercial companies and there are already several products available in this area: Quick Draw With Google[2], an interactive drawing recognition tool recently launched by Google, Google Handwriting Input in 82 Languages[3] and Microsoft Handwriting Input for Windows[4] to name just a few of them. These OCR systems are in general agnostic on the document structure and content they process.

We faced the need for Intelligent Character Recognition (ICR) a few years ago, but recently a precise requirement for it emerged: the *National Institute of Diabetes, Nutrition and Metabolic Diseases "Prof. N.C. Paulescu"* (NIDNMD) is the owner of the world oldest medical records in diabetes, still active in a city (Bucharest), from 1941 until today (Ionescu-Tîrgoviște, 2001). The archive is paper-based, containing more than 220,000 medical records. The documents are structured, according to three types of forms, but the filling in of the blanks is handwritten. Most fields which medical experts are interested in are numerical ones (i.e. age, height, weight, diabetes type, blood glucose, etc.). Knowing what to expect from the document structure makes the handwritten recognition task a little bit easier to accomplish and enables automatic validation of data consistency.

Thus, we engaged, together with the medical experts from NIDNMD, into transforming the paper-based archive into a digital searchable archive. This project, called InsyderPal, is still under the last phase of evaluation, within the national program PNIII (*Fundamental and Frontier Research-PN-III-P4-IDPCCF-2016*), meant to support multi-, inter- and trans-disciplinary research.

[2] https://quickdraw.withgoogle.com/

[3] https://research.googleblog.com/2015/04/google-handwriting-input-in-82.html

[4] https://www.howtogeek.com/297443/how-to-use-handwriting-input-on-windows-10/

Provided it is funded, the project is expected to have also a standardization effect for the digital recording of the diabetes data in the specialized medical institutions, all over the country.

RACAI has previously worked with sequence based processing algorithms. The ICR tool, recently implemented at RACAI and which will be further developed in this project, is based on recurrent neural networks with a connectionist temporal classification layer (Graves and Schmidhuber, 2009). Our proposed ICR system architecture is composed of 3 main tools, two already prototyped: (a) a segmentation tool for text lines extraction using vertical projection-profile (Likforman-Sulem et al., 2007) (b) a classification and document structure understanding (DSU) tool, and (c) the ICR recognition tool itself. While the first (a) and later mentioned (c) tools already exist, the document understanding tool will be later developed in the project. Its primary purpose is to support the process of automatic classification and detection structures within documents and to enable the creation of data sanity checks (e.g. the overall cholesterol must be equal with the sum of HDL and LDL cholesterols). Depending on the output of the DSU system, ICR will be performed either by CNNs whenever the input is numerical and easily separable or by using bidirectional LSTMs when we are dealing with continuous handwritten data.

To train the ICR tool, a collection of learning examples has to be compiled. An example is a triple formed by an image segment (extracted by means of the segmentation tool mentioned above), a label identifying the region in the scanned image of a record from where the segment was extracted (according to the templates mentioned at step b) above) and the correct result of its recognition (the truth). We estimate that for a high precision ICR process, the training set should include around 100,000 examples. The training data creation is an interactive process, requiring a lot of human involvement (especially in defining the truth for an image segment, but not only).

The project also has as an objective the implementation of an application allowing the professionals to update and exploit the digitalized archive as well as to extract relevant information (in Romanian language) from specialized literature.

There are several language resources backing-up the information extraction, including a heavily annotated and carefully validated domain specific BioRo corpus (currently 561,978 sentences, with almost 10 million tokens, MSD tagged, lemmatized, NERC marked-up), a thematic term dictionary (about 8,000 entries) and a collection of pre-processing tools (tokenizer, tagger, lemmatizer, NER) trained for this specific medical language (Mitrofan and Tufiș, 2016), (Mitrofan, 2017), (Mitrofan and Tufiș, 2018). All these resources will be significantly extended, both quantitatively and qualitatively.

6. Turning Old Romanian Documents Written with Cyrillic Alphabets into a Latin Character Set Electronic Corpus

An even more challenging project requiring ICR is "CyRo", a typical Digital Humanities project, the objective of which is to create a prototype processing chain, turning old Romanian documents written with Cyrillic characters into editable documents and then transliterating them into Latin alphabet. This is also a project proposal, still under the last phase of evaluation, within the national program *Fundamental and Frontier Research-PN-III-P4-IDPCCF-2016*.

The consortium of this project is a large one with experts in language, image processing and deep learning technology, from the Institute of Computer Science of the Romanian Academy in Iași, "A.I. Cuza" University of Iași, the University of Bucharest and the Research Institute for Artificial Intelligence of the Romanian Academy in Bucharest, experts in old and rare manuscripts from the Library of the Romanian Academy and experts in diachronic Romanian and Slavonic languages from "Al. Philippide" Philology Institute of the Romanian Academy in Iași, the Faculty of Letters of the University of Bucharest and the Faculty of Letters of "A. I. Cuza" University of Iași.

While the Cyrillic into Latin text transliteration is an almost error-free process (Cojocaru *et al.*, 2016), the main difficulties are in transforming scanned old documents into editable texts. Cojocaru and her colleagues (2016) reported very good results on implementing an integrated processing flow meant to OCR and transliterate in the Latin alphabet recent Romanian prints (1951-1989) written with Cyrillic characters in the former Moldavian Soviet Socialist Republic.

Yet, the CyRo project faces more difficulties: three different Cyrillic alphabets used during the targeted period, plus two types of handwritten documents: the so-called semi-uncial handwriting and regular cursive handwriting. In semi-uncial handwriting each drawn character imitates a printed character and is separated from the neighboring ones. Therefore this type of handwriting is supposed to be easier to process than cursive writing, which includes ligatures, overwriting, the use of multiple phonetic values for some letters, where the designation of proper names in many cases appears with initial lowercase, where abbreviations and diacritics without phonetic values in Romanian are frequently used, where there are combinations between abbreviations and overwriting, inconstant denotation of numerals, and sometimes spaces separating words are missing (*scriptio continua*).

Another important group of difficulties (Cristea et al. 2012), (Simionescu et al., 2012) is triggered by the diachronic changes of the Romanian language (phonetic, morphological, lexical, syntactic). Although, for realistic promises, we explicitly address only documents printed and with semi-uncial handwriting, we will also aim to do extensive experiments on cursive handwriting, but offering also the expert the possibility to make corrections on the final interpreted document.

The ICR technology created in CyRo will facilitate the development of a quasi-exhaustive diachronic corpus of the Romanian language, which will be used in applications and studies supposed to have a high impact on the study of the Romanian language, as well as, more generally, on the family of Romance languages. Moreover, the new methods and tools laid down in CyRo are expected to inspire a whole generation of interpretative instruments adapted to Slavic languages, for the automatic interpretation of old documents written in Cyrillic.

7. Conclusions

Some of the projects, approaching the end, mentioned in this article are described in larger details in dedicated papers (Barbu Mititelu *et al.*, 2018; Mitrofan and Tufiș, 2018). We tried to offer a global view of the relevant projects we and our colleagues carry on in the area of AI, NLP and DH domains. Unlike many of our previous projects, the current ones are nation-wide, bringing together more R&D groups and aiming at more ambitious targets. Some results will be transferred to the industry for further development and industrialization and their distribution will be restricted. However, most outcomes of the mentioned projects will be freely accessible to the interested public. Several datasets (lexicons, treebank, text and speech corpora, word embeddings) created for training different machine learning modules and tools (TTL, SSPR, MLPLA) are already available and others (ASR, TTS, KWS, ICR) will be made public (ensuring anonymisation, whenever necessary, of the personal information, e.g. in the processed medical records and the speech recordings), as soon as they are ready.

8. Acknowledgements

The people who are involved in the above mentioned projects are too many to be listed here, but we want to express our gratitude to all of them (data providers, researchers, PhD and MSc students), for their enthusiastic past, present and future engagement and hard work.

We also thank to the funding bodies that made these endeavors possible: Romanian Academy, Romanian Ministry of Education and Research, UEFISCDI (Executive Agency for Higher Education, Research, Development and Innovation Funding), Humboldt Foundation and Institute for German Language IDS-Mannheim.

9. Bibliographical References

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), Dubrovnik, Croatia, 29 September – 1 October 2016

Barbu Mititelu, V., Tufiș, D., Irimia, E. (2018): The Reference Corpus of the Contemporary Romanian Language (CoRoLa). Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaky, Japan.

Bohnet, B. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Boroș, T. and Dumitrescu, S. D. (2017). A "small-data"-driven approach to dialogue systems for natural language human computer interaction. In *The 9th Conference on Speech Technology and Human-Computer Dialogue (SPED 2017)*. Bucharest, Romania

Boroș, T., and Dumitrescu, S. D. (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 EcoSystems*, pp. 98-102, ACM.

Boroș, T., Ion, R., & Tufiș, D. (2013). Large tagset labelling using Feed Forward Neural Networks. Case study on Romanian Language. In *Proceedings of ACL*, Sofia, pp. 692-700.

Cojocaru, S., Colesnicov, A., Malahov, L. (2017). Digitization of Old Romanian Texts Printed in the Cyrillic Script. In *Proceedings of DATECH 2017*, 1-2 June, Göttingen

Cosma, R., Cristea, D., Kupietz, M., Tufiș, D., Witt, A. (2016). DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora. In Bański, P., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., Kupietz, M., Lüngen, H., and A. Witt (eds.) The 4th Workshop on Challenges in the Management of Large Corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož / Paris: ELRA, pp. 28-32.

Cristea, D., Simionescu, R., & Haja, G. (2012). Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations. In *LREC* (pp. 923-927).

Cucu, H., Buzo, A., Besacier, L., & Burileanu, C. (2014). SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian. *Speech Communication*, *56*, 195-212.

Dumitrescu, Ș. D., Boroș, T. and Tufiș, D. (2017). RACAI's natural language processing pipeline for Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. ACL Vancouver, Canada, pp. 174-181.

Graves, A., & Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems* (pp. 545-552).

Ion, R., Irimia, E., Barbu Mititelu, Verginica. (2018). Ensemble Romanian Dependency Parsing with Neural Networks. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaky, Japan.

Ionescu-Tîrgoviște, C., *Diabetul in Romania* (*Analele Academice Diabetologice)*,(2001) Briliant, Bucuresti, 2001.

Kupietz, M., Andreas 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 *Proceedings of the 5th* CMLC, Birmingham, pp.15-20.

Lei, T., Xin, Y., Zhang, Y., Barzilay, R., Jaakkola, T. (2014). Low-Rank Tensors for Scoring Dependency Structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL 2014), June 23-25, Baltimore, Maryland, USA.

Likforman-Sulem, L., Zahour, A., Taconet, B. (2007). Text line segmentation of historical documents: a survey. In *International Journal of Document Analysis and Recognition*, Vol 9(2-4), pp.123-138.

Mitrofan, M. (2017). Bootstrapping a Romanian Corpus for Medical Named Entity Recognition. In *Proceedings of RANLP,* Varna, Bulgaria, pp. 501–509.

Mitrofan M., and Tufiș, M.(2016). Building and evaluating the Romanian medical corpus. In Proceedings of the 12th International Conference ”Linguistic Resources and tools for processing the Romanian language”, pp. 29–36.

Mitrofan M., and Tufiș, D. (2018). BioRo: The Biomedical Corpus for the Romanian Language. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaky, Japan.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2), pp. 95-135.

Păiș, V., and Tufiș, D. (2018). Computing Distributed Representations of Words using the CoRoLa Corpus. Proceedings of the Romanian Academy, Series A, vol 19. No.1/2018

Simionescu, R., Cristea, D., & Haja, G. (2012). Inferring diachronic morphology using the Romanian Thesaurus Dictionary. In *Proceedings of the 8th International Conference “CONSILR”, April, Bucharest* (pp. 85-92).

Stan, A., Watts, O., Mamiya, Y., Giurgiu, M., Clark, R. A., Yamagishi, J., & King, S. (2013, August). TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. In *INTERSPEECH* (pp. 2331-2335).

Tufiș, D., & Ștefănescu, D. (2012). Experiments with a differential semantics annotation for WordNet 3.0. *Decision Support Systems*, *53*(4), 695-703.

Tufiș D., Barbu Mititelu V., Irimia E., Dumitrescu Ș. D., Boroș T. (2016). The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), European Language Resources Association (ELRA), ISBN: 978-2-9517408-9-1, pp. 2516-2521, Portorož, Slovenia

ANNEXES

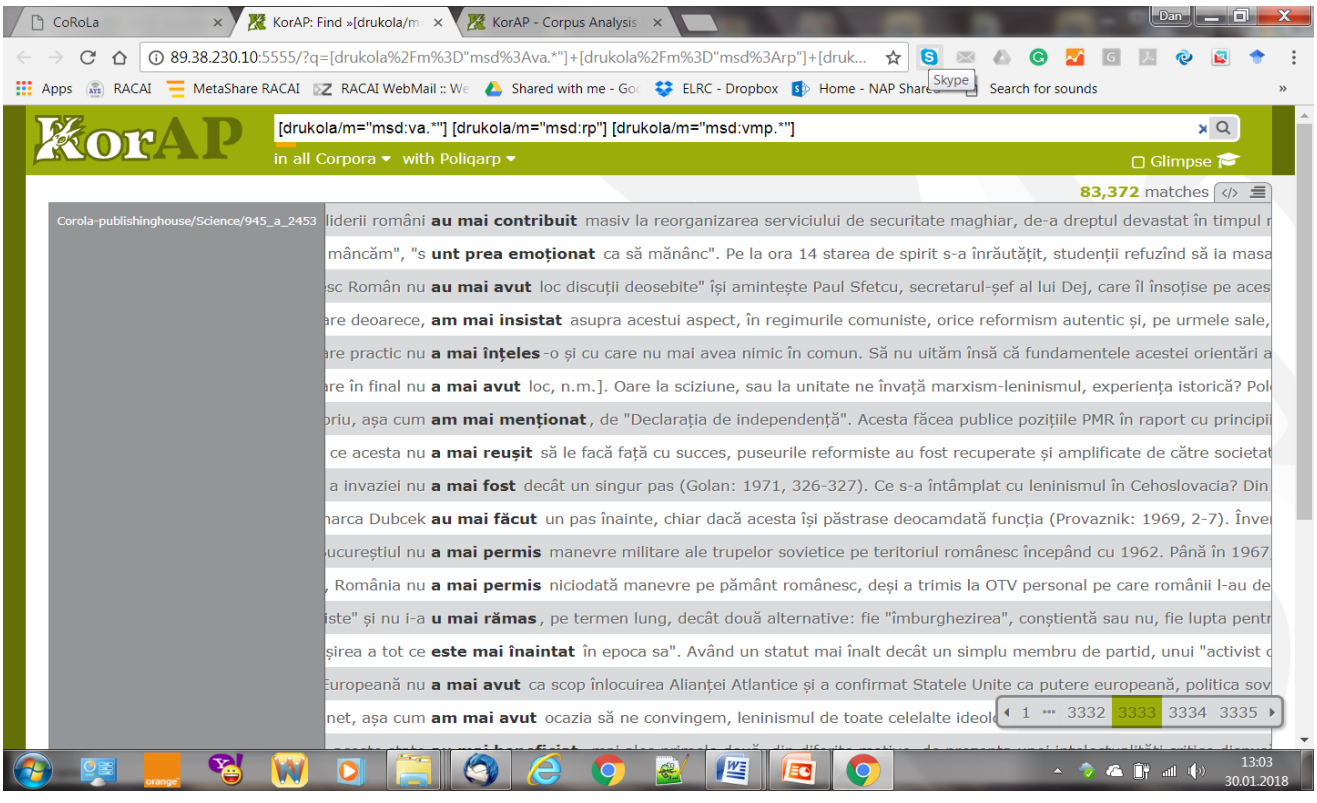

Figure 1: KoRaP Search Interface for CoRoLa
Query (Poliqarp): an auxiliary verb followed by an adverb followed by a participle

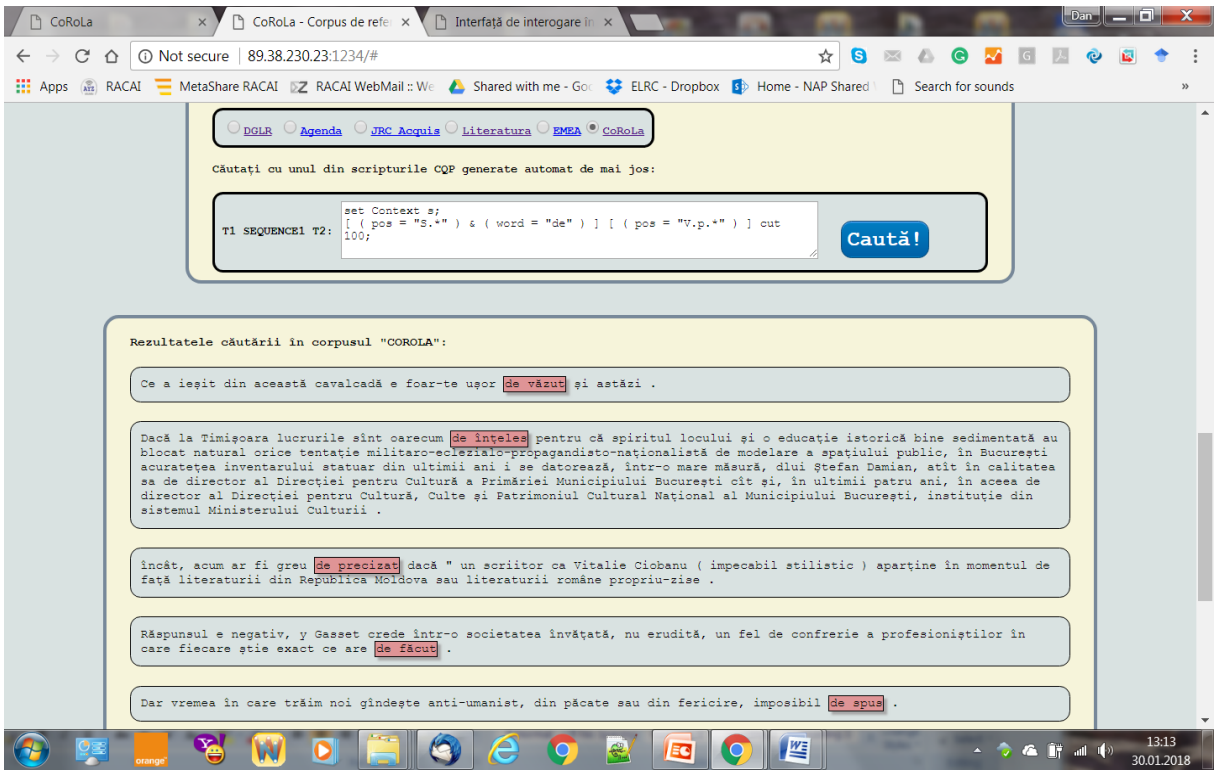

Figure 2: NLP2CQP Search Interface for CoRoLa
Query (NL language): „100 de fraze în care prepoziția "de" este urmată imediat de un verb la participiu = 100 sentences in which the preposition "de" is immediately followed by a participle verb"

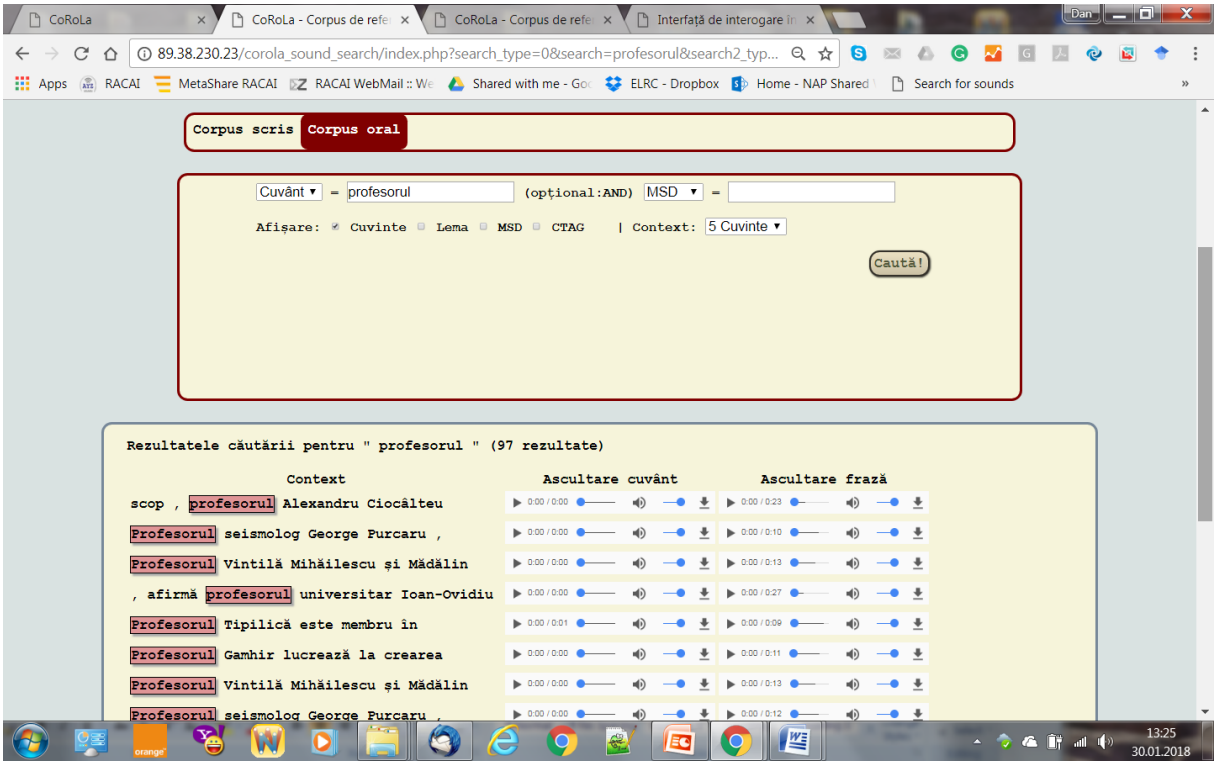

Figure 3: OCQP Search Interface for CoRoLa
Query: oral fragments containing the wordform „profesorul" (the professor)