The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language

Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia, Ştefan Daniel Dumitrescu, Tiberiu Boroş
Research Institute for Artificial Intelligence “Mihai Drăgănescu”
13, Calea 13 Septembrie, 050711, Bucharest, Romania
{tufis, vergi, elena, sdumitrescu, tibi}@racai.ro

Abstract

The article describes the current status of a large national project, CoRoLa, aiming at building a reference corpus for the contemporary Romanian language. Unlike many other national corpora, CoRoLa contains only - IPR cleared texts and speech data, obtained from some of the country’s most representative publishing houses, broadcasting agencies, editorial offices, newspapers and popular bloggers. For the written component 500 million tokens are targeted and for the oral one 300 hours of recordings. The choice of texts is done according to their functional style, domain and subdomain, also with an eye to the international practice. A metadata file (following the CMDI model) is associated to each text file. Collected texts are cleaned and transformed in a format compatible with the tools for automatic processing (segmentation, tokenization, lemmatization, part-of-speech tagging). The paper also presents up-to-date statistics about the structure of the corpus almost two years before its official launching. The corpus will be freely available for searching. Users will be able to download the results of their searches and those original files when not against stipulations in the protocols we have with text providers.

Keywords: corpus annotation, corpus design, IPR-clearing, metadata reference corpus, Romanian language

1. Introduction

One of the most solicited resources out of our MetaNet4U distributions via Meta-Share platform (http://ws.racai.ro:9191/) is ROMBAC - the Romanian Balanced Corpus (Ion et al., 2012) containing 44,117,360 tokens covering four domains (News, Medical, Legal, Biographic and Fiction) and built around the earlier RoCo_News corpus (Tufiş and Irimia, 2006). The great feedback we received from the user community, as well as their growing need of larger and still balanced corpora encouraged us to start a more ambitious project, namely a reference corpus for contemporary Romanian, about 15 times larger, including a speech component and a treebank.

The new project, called CoRoLa (**C**orpus of C**o**ntemporary **Ro**manian **La**nguage), started in 2014 with a first version planned to be opened for the public at the end of 2017. The project, rated as a priority project of the Romanian Academy, has been joined by the Institute for Computer Science in Iaşi. Besides the two institutions which were commissioned by the Romanian Academy to run the CoRoLa project, the corpus developments is voluntarily contributed by linguist experts from the Linguistic Institute “Al. Philippide” of Iași and many Master and PhD students from University “A.I. Cuza” of Iași, University “Politehnica” of Bucharest and the University of Bucharest.

Contemporary Romanian Language is the last phase in the evolution of the Romanian language, starting, according to the specialists, after the Second World War. Although initially planned to cover the period 1945-present, with two sub-periods (1945-1990, 1990-present), with clear differences mainly at the lexical level, that would have given this corpus a definite diachronic dimension, technical difficulties will restrict the period represented in CoRoLa to 1990s-present:while for the last couple of decades there is an important amount of electronic texts available, this is not the case for the texts from the period 1945-1990, for which considerable effort needs to be invested in digitizing the texts (scanning, OCRizing and correcting them), as well as in IPR-cleaning.

2. Structure of the corpus

In its first public version, CoRoLa will contain more than 500 million tokens and more than 300 hours of transcribed speech and it will be IPR cleared.

All functional styles will be represented: scientific, official (administrative + juridical), journalistic, memorialistic and imaginative and the corpus is supposed to be representative for the literary language.

The colloquial style is not a major concern for us, because its processing leads to poor results. Nevertheless, it will be represented in the corpus as it is used in imaginative writing.

The provisional structure of CoRoLA is detailed in (Barbu Mititelu and Irimia, 2014). In designing the structure of the corpus we draw upon the composition of many different national reference corpora: British National Corpus (http://www.natcorp.ox.ac.uk/ corpus/index.xml), Russian National Corpus (http://www.ruscorpora.ru/en/), Czech National Corpus (Čermák & Schmiedtová, 2003), the Reference Corpus of the Contemporary Portuguese (http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc), Polish National Corpus (Przepiórkowski et al., 2011), Bulgarian National Corpus (Koeva et al., 2012), Croatian National Corpus (http://www.hnk.ffzg.hr/ struktura_en.html), International Corpus of Arabic (http://www.bibalex.org/unl/frontend/Project.aspx?id=9).

In (Tufiş, 2015) we presented the stratified sampling strategy (Biber, 1993; Passonneau et al., 2014), aimed at balancing the disparity between quantities of available data for each functional style and ensure the representativeness of the corpus.

The corpus covers 5 large domains (arts &culture, society, science, nature and others) which are further refined into 71 sub-domains. The domains and sub-domains classification is based on the Wikipedia one, with refinements justified sometimes by the fine granularity of Wikipedia and some other times by the impossibility of finding enough texts for either relatively new or narrow sub-domains. The corpus now covers the following 51 sub-domains grouped into domains as detailed below. Each domain has a set of texts that are attributed the sub-domain "other", as none of those we specified is appropriate for classifying those texts.

- Domain: arts &culture. Sub-domains: literature, art history, folklore, film, architecture, painting/drawing, design, theatre;
- Domain: society. Sub-domains: politics, law, administration, economy, health, sports, gossip, social events, education, tourism, religion, entertainment;
- Domain: nature. Sub-domains: environment, natural disasters, universe, natural resources;
- Domain: science. Sub-domains: mathematics, informatics, medicine, archaeology, engineering, technics/technology, agronomy, constructions, pharmacology, enology, pedagogy, geography, economy, history, psychology, sociology, anthropology, religious studies and theology, juridical sciences, linguistics, political sciences, philosophy, philology, biology, physics, astronomy, chemistry.

CoRoLa also includes a syntactically annotated sub-corpus (treebank) and an oral component. The former has 9500 sentences chosen from various domains and functional styles, in an effort of preserving the balanced character of CoRoLa. It is annotated within the dependency grammar framework, following the guidelines within the Universal Dependency project (http://universaldependencies.github.io/docs/u/overview/syntax.html) and using a set of syntactic relations based on the one within the same project (http://universaldependencies.github.io/docs/u/dep/index.html), but adjusted so that to capture Romanian specificities at the syntactic level.

The oral data (targeted: at least 300 hours of transcribed recorded speech) is accompanied by annotations on: speech segmentation at sentence level, pauses, non-lexical sounds, grapheme-phoneme alignments and explicit marking of the accent.

All current textual data is morpho-lexically processed (tokenized, POS-tagged and lemmatized). The morpho-syntactic annotations of the textual data are provided in-line while possible further layers of linguistic annotation (especially at the discourse level) for textual data and specific annotation for speech data will necessitate a mix of stand-off and in-line markup.

All the textual documents collected so far (95,348 files, see Section 4) are accompanied by standardized metadata, conformant with the CMDI model (Component Meta Data Infrastructure). The metadata are created either automatically or manually. The former method is used for:

1. texts crawled from the web (usually newspaper articles or blog posts), with an intermediary stage of mapping the classifications of domains from various websites to ours.
2. 14,294 files (representing the ROMBAC seed corpus of the larger CoRoLA corpus): metadata created for the MetaShare platform was specified as XCES header (inside the files). As XCES headers contain a superset of CMDI, generating the CMDI stand-off representation starting from them was a simple task.

The manual method is used for the rest of texts. In order to ease work, a tool for text cleaning and metadata completion was developed (Moruz & Scutelnicu, 2014). The metadata annotators are provided with a detailed instruction manual, thus ensuring uniform treatment of various cases.

3. Data Collection and Cleaning

The resource we are building has two important features: it is representative for the contemporary language, covering all literary language registers and styles and it is IPR cleared, which is a challenging task, given the modern IPR-driven society. The categories of content excepted from the IPR restrictions in Romania are: political, legislative, administrative and juridical. For all the other types of content, to ensure the volume and quality of the data in the corpus, as well as to clear the IPR on these data, our endeavour was to establish collaborations with publishing houses and editorial offices. So far, we have signed agreements with major publishing houses, magazines and newspapers [1] . Additionally, four representative bloggers[2] have agreed to allow us to include some of their posts in the corpus. Oral texts (read news, live transmissions and live interviews) (one hour per working day) were provided by RADOR, the press agency of Radio Romania (http://www.rador.ro/).

Cleaning the data received from our providers and converting it to an adequate format for our pre-processing tools assumes significant work, which to a large extent was automated (Moruz & Scutelnicu, 2014, the same tool involved in metadata creation mentioned in Section 2): the text is extracted from the PDF files, paragraph limits are recuperated, column marking newlines, as well as hyphens at the end of the lines are erased. Still, a lot of manual work remains to be done: separating articles from periodicals in different files, removal of headers, footers,

[1] publishing houses: Humanitas, Polirom, Romanian Academy Publishing House, Bucharest University Press, "Editura Economică", ADENIUM Publishing House, DOXOLOGIA Publishing House, the European Institute Publishing House, GAMA Publishing House, PIM Publishing House; magazines and newspapers: România literară, DCNEWS, Muzica, Actualitatea muzicală, Destine literare, the school magazine of Unirea National College from Focşani.

[2] Simona Tache , Dragoș Bucurenci , Irina Șubredu and Teodora Forăscu

page numbers, figures, tables, dealing with foot- or end-notes, with text fragments in foreign languages, with excerpts from other authors, etc. When extracted from its original source, the textual content is converted into the UTF-8 encoding and saved as plain text documents. Another issue is related with the use of diacritics. Although we consider only texts written with diacritics (the quality of the linguistic annotation would be badly affected in their absence), we have to make sure that the correct encodings are used throughout the corpus. This problem is generated by a long time co-existence of two sets of codes for the Romanian diacritics. The affected characters are ș (&scomma;) and ț (&tcomma;), which before the current standardization were written as (&scedil;) and (&tcedil;).

The original texts received from our partners or crawled from the web are kept separately on our servers. However, access to them is restricted: protocols signed with text providers limit our rights to that of cleaning and processing, as well as allowing third party's access to the snippets returned as a result of processing their queries. For texts without IPR restrictions, access to the original can be offered on request.

CoRoLa is currently exploitable using the IMS Open Corpus Workbench, an open source medium (CWB, http://cwb.sourceforge.net/) that allows complex searching with multiple criteria and support for regular expressions (regex). Additionally, an interface allows the regex inexperienced users to formulate their queries in constrained Romanian language, which are translated into CQP queries. CWB has been installed and coupled with our processing chain which produces the adequate annotated format for morphological and shallow syntactic search criteria.

Within a newly launched project, in collaboration with IDS Mannheim, the CoRoLa corpus will be added to a larger multilingual collection of corpora and, in the near future, we will switch to the KorAP corpus management platform (Bańksi et al., 2014; Cosma et al., 2016).

4. Current CoRoLa Content

As mentioned before, the CoRoLa corpus includes both textual and speech data. The subsections below will focus on each of them in turn.

4.1 Textual Data

At the moment, the corpus contains more than 140 million of words (excluding punctuation) distributed over the 51 out of the 71 sub-domains mentioned in Section 1. In Table 1, we provide a coarse-grain domain classification of the texts, as well as quantitative data (number of lexical tokens, including punctuation) related to each domain.

The TTL (Ion, 2007) chain ensures, at the time of this writing, the following specific functionalities: sentence splitting, tokenisation, tiered-tagging (Tufiș, 1999), lemmatising and chunking. The tagset (MSD-tags) is compliant with the MULTEXT-EAST specifications (Erjavec, 2012). It is already trained to deal with Romanian, English and French. The output is provided in two formats: as a vertical text with tagging, lemmatization and chunking information on the same line with the corresponding token or as an xml encoded (XCES) document. The average accuracy of the processing flow is about 97.5% (Tufiș et al., 2008). The vertical text format is used as input to the dependency parser (see below).

Table 1. Domain and style distribution of textual data

Domain	Tokens	Style	tokens
arts&culture	50,971,951	journalistic	57,478,242
society	48,187,517	science	59,114,059
science	46,220,700	imaginative	29,299,032
nature	946,928	memoirs	23,204,286
others	45,206,113	administrative	1,644,783
		law	18,321,366
		others	2,471,441
TOTAL	191,533,209	TOTAL	191,533,209

Table 2 shows an example of the tabular representation of the morpho-syntactic processing of a Romanian sentence, as produced by the TTL NLP chain. The Syn-chunk column shows for each lexical item the syntactic chunk it belongs to. The chunks may be embedded (the outermost chunk is the one at the left extreme of the label). For instance, in Table 2, the Syn-chunk label of the word "textuală","Np#2,Ap#1", should be read as: "the word textual is an adjectival phrase modifying the head noun "prelucrare" (Np#2). The Np#2 covers the sequence "prelucrarea textuală a datelor".

Table 2. Example of the tabular output of TTL

Wordform	Lemma	MSD-tag	Syn-Chunk
Acest	acest	Dd3msr---e	Np#1
exemplu	exemplu	Ncms-n	Np#1
ilustrează	ilustra	Vmip3	Vp#1
prelucrarea	prelucrare	Ncfsry	Np#2
textuală	textual	Afpfsrn	Np#2,Ap#1
a	al	Tsfs	Np#2
datelor	dată	Ncfpoy	Np#2
.	.	PERIOD	

As mentioned in section 2, the first release of CoRoLa corpus will include the Romanian UD-compliant treebank as well. Table 3 shows an example of dependency parsing (CoNLL format, ignoring the PHEAD and PDEPREL columns) of the same sentence exemplified in Table 2.

Table 3. Example of the tabular dependency parse

ID	Wordform	Lemma	MSD-tag	HEAD	DEPREL
1	Acest	acest	Dd3msr---e	2	amod
2	exemplu	exemplu	Ncms-n	3	subj
3	ilustrează	ilustra	Vmip3	0	ROOT
4	prelucrarea	prelucrare	Ncfsry	3	dobj
5	datelor	dată	Ncfpoy	4	nmod
6	textuale	textual	Afpfp-n	5	amod
7	.	.	PERIOD	3	punct

4.2 Speech Data

The interviews and news recordings are accompanied by transcriptions (observing the current orthography). Out of the currently collected data (more than 135 hours of transcribed speech), about 37% was automatically pre-processed and the transcriptions were XML encoded with mark-up for lemma, part-of-speech and syllabification. The time alignments of the words and their phonemes have also been automatically encoded in separate files. The processed speech data serves two projects pursued in parallel: CoRoLa and ANVSIB (http://speed.pub.ro/anvsib). One major concern for the ANVSIB project is to create speech language models that fit into the memory of mobile devices. Boroș and Dumitrescu (2015) describe the experiments performed in modelling the major processing steps for text-to-speech production: syllabification, phonetic transcription, POS tagging, and lexical stress prediction. The approach used for modelling and speech synthesis is based on deep neural networks (DNN) with 2 or 3 hidden layers and a varying number of input and output neurons. In their paper they describe how different labelling and feature encoding strategies increase the performance of the DNN classifier on the target task. As compared to the previous MIRA-based tools (Boroș et al., 2013), the new models are much smaller (between 32 and 539 times), with acceptable loss of processing accuracies (around 1.5%), with acceptable processing accuracies (see Table 4). One mention is necessary for the POS tagging evaluation: the new DNN model tries to solve two problems at once: the proper POS tagging (611 POS tags) and the NER (76 labels). While the proper POS tagging is significantly better than reported global result (almost 98%), the NEs labelling is rather poor. Previously, this labelling in the training corpus was achieved by a post-tagging rule-based recognizer. The number of NEs in the training corpus was not enough for a supervised training of the DNN models, so solving the NER within the same step or as a separate step is subject to future experimentation. Anyway, considering the current accuracies and the much faster runtime speed, we think that the DNN solution is the right solution for the large quantity of speech data we already collected and which will be further extended, according to our conventions (at least 200 more hours of recorded speech).

Table 4. Sizes and accuracy losses for our previous and current speech models

TASK	Size and accuracy MIRA-based models	Size and accuracy DNN-based models	Accuracy loss
Syllabification	9426.5 KB/ 99.01%	36.7 KB/ 98.23%	0,78%
Phonetic transcription	1389.1 KB / 96.29%	43.4 KB/ 96.16%	0,13%
POS tagging	98.19% 96 MB	95.16% 178 kB	3,03%
Lexical stress prediction	6 MB/ 98.80%	110 KB 97.67%	1,13%

The speech data which we collected until now, partly processed as shown in Table 6, are the following (see also Table 5):

- RASC (Romanian Anonymous Speech Corpus) is a crowd-sourcing initiative to record a sample of sentences randomly extracted from Ro-Wikipedia as described in (Dumitrescu et al., 2014). The corpus is aligned at phoneme/word level.
- RSS-ToBI (Romanian Speech Synthesis Corpus) is a collection of high quality recordings compiled by (Stan et al., 2011) and designed for speech synthesis. It was enhanced with a prosodic ToBI-like (Tone and Break Indices) annotation (Boroș et al., 2014). It is aligned at phoneme/word level.
- RADOR (Radio Romania) is a collection of radio news and interviews, provided daily by the Romanian Society for Broadcasting. At the time of this writing, the transcriptions are under processing. They are partially aligned at phoneme/word level. During the processing of the audio files, the musical sequences, and the commercials were eliminated. The sequences containing overlapped speech or not intelligible were also eliminated.
- The fourth speech sub-corpus is produced by professional speakers reading a collection of sentences containing most of the interesting phonetic structures in Romanian. The recording has been done in an acoustic room.

A useful search utility for the speech corpus has been lately developed. During the speech transcription, each word is marked with the beginning and finish times (the resolution is 50ms) so that the user may retrieve and listen to the sentences containing the searched words.

Table 5. Speech corpora already collected

Corpus	Type	Source	Time length (h:m:s)
RASC	Many speakers	RoWikipedia	04:22:02
RSS-ToBI	Single speaker	News & Fairy tales	03:44:00
RADOR	Many speakers	News& interviews	106:52:33 (out of which processed 37:51:23)
Acoustic room recordings	Two speakers (male, female)	Selected sentences	19:58:46
TOTAL			134:57:24

Table 6.Currently pre-processed speech corpora

Corpus	sentences	words	phonemes
RASC	2,866	39,489	270,591
RSS-ToBI	3,500	39,041	235,150
RADOR	10,349	360,556	2,304,673
	16,715	439,086	2,810,414

5. Conclusions

Although large amounts of texts are out there on the web, creating an IPR clear reference corpus is quite a challenge. On the one hand, it implies vast efforts invested in persuading IPR holders to contribute to a cultural action in a way that does not hinder their marketing plans. On the other hand, it means reaching agreement on what texts and how much of them to include in the corpus. The CoRoLa structure (text types and target quantities) of the language data (Barbu Mititelu & Irimia, 2014) was built according to what our collaborators, IPR holders, could offer us and to international practices in corpus design.

However, the large quantities of web data previously collected are not discarded: they are stored and used in training specialized statistical models supporting different data-driven applications (CLIR, Q&A, Sentiment analysis, SMT, ASR, and TTS).

So far, the CoRoLa development has been progressing as planned, reaching almost half of the targeted size. Our national project has two more years to go for the publicly opening of the first operational version of the corpus. CoRoLa is and will continue to be automatically annotated, but a fragment of it (about 1 million tokens) will be manually validated. The corpus will be available for search for all those interested in the study or processing of the Romanian language.

We consider that our approach for building and exploiting the corpus can serve as a model for other corpus developers: we show that IPR holders can be convinced to join our efforts of creating quality corpora.

6. Acknowledgements

Although this article reflects mainly ICIA's point of view, the development of the CoRoLa corpus is carried out in close cooperation with the Institute of Theoretical Computer Science from Iași, partnership which is highly appreciated.

Our work could not have been possible with similar results without the collaboration of our partners, who kindly agreed to offer us access to (oral and written) texts.

7. 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., Irimia, E. (2014). The Provisional Structure of the Reference Corpus of the Contemporary Romanian Language (CoRoLa). In M. Colhon, A. Iftene, V. Barbu-Mititelu, D. Tufiș (eds.) *Proceedings of the 10th Intl. Conference "Linguistic Resources and Tools for Processing Romanian Language"*, University of Craiova, pp. 57—66.

Barbu Mititelu, V., Mărănduc, C., Irimia E. (2015). Universal and Language-specific Dependency Relations for Analysing Romanian. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden, August pp. 28–37.

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

Boroş, T., Dumitrescu, Ș. (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 (MEDES'15)*, 25-29 October 2015, Caraguatatuba/Sao Paulo, Brazil

Boroş, T., Stan, A., Watts, O., Dumitrescu, S.D. (2014). RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus. In *Proceedings of 9th LREC 2014, Reykjavik*, Iceland, 25-31 May, European Language Resources Association (ELRA).

Cosma, R., Cristea, D., Kupietz· M., Tufiș, 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).

Cristea, D., Pistol I.C. (2012). Multilingual Linguistic Workflows. In Cristina Vertan and Walther v. Hahn (Eds.) *Multilingual Processing in Eastern and Southern EU Languages. Low-resourced Technologies and Translation*, Cambridge Scholars Publishing, UK.

Dumitrescu, S. D., Boroş, T., Ion, R. (2014). Crowd-Sourced, Automatic Speech-Corpora Collection – Building the Romanian Anonymous Speech Corpus. CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era,

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.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3), pp. 380--409.

Ion, R. (2007). Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis in Computer Science. Romanian Academy (in Romanian).

Ion, R., Irimia, E., Ştefănescu, D., Tufiș. D. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In Nicoletta Calzolari et al. (Eds.) *Proceedings of the 8th LREC, Istanbul, Turkey*, European Language Resources Association (ELRA).

Irimia, E., Barbu Mititelu V. (2015). Building a Romanian Dependency Treebank, In *Proceedings of Corpus Linguistics 2015*, Lancaster University, UK, pp. 21--24

Mărănduc C., Perez, A.C. (2015) A Romanian dependency treebank, In *Proceedings of CICLing 2015*, Cairo.

Moruz, A., Scutelnicu, A. (2014). An Automatic System for Improving Boilerplate Removal for Romanian Texts, in M. Colhon, A. Iftene, V. Barbu Mititelu, D.

Cristea, D. Tufiș (eds.) *Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language"*, Craiova, 18-19 September 2014, Editura Universității „Alexandru Ioan Cuza", Iași, pp. 163--170.

Passonneau, R.J., Ide, N., Su, S., Stuart, J. (2014) Biber Redux: Reconsidering Dimensions of Variation in American English. In *Proceedings of COLING 2014*, pp. 565--576.

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), pp. 442--450.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources*.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (eds) *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28--33

Tufiş, D. and Irimia, E. (2006). RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC Conference*. Genoa, Italy, May 2006, pp. 869--872.

Tufiş, 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, May 2008, European Language Resources Association (ELRA).

Tufiş, D., Boroș, T. (2015) - Challenges in building a publicly available reference corpus. *Tutorial at EUROLAN 2015 Summer School*, Sibiu, August 13-25, 2015.