Published in: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente et al. (Hg.) (2016): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 28-32

DRuKoLA - Towards Contrastive German-Romanian Research based on Comparable Corpora

Ruxandra Cosma[1], Dan Cristea[2,3], Marc Kupietz[4], Dan Tufiş[5], Andreas Witt[4,6]

[1]University of Bucharest, Faculty of Foreign Languages
[2] *Alexandru Ioan Cuza* University of Iaşi, Department of Computer Science
[3] Romanian Academy, Institute for Computer Science - Iaşi
[4] Institut für Deutsche Sprache, Mannheim
[5] Institute for Artificial Intelligence *Mihai Drăgănescu*, Bucharest
[6] Heidelberg University, Department of Computational Linguistics

ruxandracosma@gmail.com, dcristea@info.uaic.ro, {kupietz, witt}@ids-mannheim.de, tufis@racai.ro

Abstract

This paper introduces the recently started DRuKoLA-project that aims at providing mechanisms to flexibly draw virtual comparable corpora from the German Reference Corpus DeReKo and the Reference Corpus of Contemporary Romanian Language CoRoLa in order to use these virtual corpora as empirical basis for contrastive linguistic research.

Keywords: Reference Corpora, Comparable Corpora, Contrastive Linguistics

1. Introduction

Corpora have increasingly been used in cross-linguistic research, where, in particular, parallel corpora have been of major importance. The usefulness of parallel resources for cross-linguistic research is obvious, as they provide bi- or multilingual, ideally aligned language data that convey the same meaning, including contextual information, and can thus serve as a basis for establishing equivalence between particular entities across different languages (cf. James 1980, Chesterman 1998). On this account, parallel data have been used as an empirical basis in many contrastive studies so far. Some examples include Altenberg (1999), Hasselgård (2007), Zufferey and Cartoni (2012), Kaczmarska and Rosen (2013), where various phenomena from English and Swedish, English, Swedish and Norwegian, English and French, Polish and Czech, respectively, have been accounted for.

Recently, there has also been growing interest in developing comparable corpora (see Sharoff et al. 2013 and the workshop series Building and Using Comparable Corpora) but so far, no comparable resources are available (at least not for German and Romanian) that would allow us to conduct cross-linguistic investigations drawing on language-specific grammatical and semantic properties. The reasons for the DRuKoLA project, as will be sketched in this paper, is to see if a common building strategy can be used for a pair of reference corpora belonging to languages of two diverse families, if a common view on the management of the two corpora can be used and if the access to them can be organised with a common corpus analysis platform. Moreover, the project will investigate how comparable virtual collections (sub-corpora) can be extracted dynamically from this shared resource and how they can serve as flexible, cost-efficient and high-qualitative empirical bases for answering comparative linguistic research questions.

2. Aims of the DRuKoLA project

The DRuKoLA project[1] that is centered around the German Reference Corpus DeReKo (Kupietz, *et al.* 2010) and the Reference Corpus of Contemporary Romanian Language CoRoLa (Tufiş, *et al.* 2015) has started in January 2016 and is a cooperation between the University of Bucharest, the Institute for the German Language in Mannheim, and research institutes of the Romanian Academy in Bucharest and Iaşi. DRuKoLa is a transdisciplinary project involving corpus linguistics, computational linguistics, applied linguistics and cross-linguistic studies, applied computer science, corpus architecture and finally also research infrastructure development. Within this broad range of areas, DRuKoLA's concrete research objectives are:

1. Construction, provision and harmonization of comparable corpora in the two languages.
2. Development of criteria for building comparable virtual sub-corpora based on DeReKo and CoRoLa, the German and, respectively, the Romanian corpus, based on metadata and other possible text properties.
3. Exploration of language-specific peculiarities of the studied languages and equivalences with respect to different parameters and structures.
4. Corpus-based comparative case studies on a) markers of modality: *haben/a avea* with *zu*-infinitives and supine

[1]DRuKoLA is funded by the Alexander von Humboldt-Foundation as a Research Group Linkage Programme between the University of Bucharest and the Institute for the German Language in Mannheim, with the Institute for Artificial Intelligence *Mihai Drăgănescu* (RACAI, Bucharest) and the Institute of Computer Science (IIT, Iaşi) of the Romanian Academy as associated partners. The acronym combines central goals of the project: corpus development and contrastive linguistic analysis (*Sprachvergleich korpustechnologisch. Deutsch - Rumänisch*).

and b) (abstract) demonstratives in German and Romanian, c) general investigation of distributional semantic and syntagmatic properties of corresponding forms and structures.

5. Development of corpus technology to share the corpus, technical and research results in a common Corpus Analysis platform.
6. Building a structure that can serve as a crystallization point for other national or reference corpora with the long-term goal of building a federated, at least European, reference corpus where each corpus is still physically located at and curated by its responsible institute, but can be dynamically extracted to different comparable corpora.

We should also mention that at least the objectives 2 – 5 are planned to be carried out in parallel and in a cyclic bootstrapping fashion. That means for example that the initial naive definition of the comparable corpora and the analysis and visualization functions of the query software will be iteratively refined based on the results of the linguistic analyses. As a welcome side-effect of this procedure, we expect to acquire a good impression of to what extent the linguistic results vary with different corpus compositions and thereby an impression of reliability and generalizability of the obtained findings.

While research objective (6) is also a long-term goal, we already expect numerous synergy effects within the range of current project. First of all, we are convinced that joining national reference or national corpora virtually, with each institute still being responsible for the curation and extension of its own resources is a much more economical and sustainable approach than building multiple comparable corpora from scratch and maintaining them on a project-basis. Another aspect concerns the development and maintenance of sustainable research software that is currently carried out individually for each reference corpus. A closer collaboration in this field with joint forces has the potential of reducing the investments on infrastructure, that are always difficult in the academic context, to a fraction. In addition to such mostly economical arguments, we are convinced that bringing the (corpus-) linguistic communities of different languages together – currently still too much centered around their philologies – has on its own a large boost potential.

3. The underlying corpus resources

Starting a project like this – situated in very different moments of corpus development and architecture – is a rare opportunity, as on the one hand we are working on and witnessing the construction of the Romanian Contemporary Reference Corpus from its beginnings and, on the other hand, are working with a very advanced German reference corpus, analysis system and technology. The collection of data for German started more than 50 years ago and the exploration of principles and methods of empirical anchoring linguistic studies at the IDS in the beginning of the nineties. The project CoRoLa started only in 2014 as a project of national priority of the Romanian Academy. The corpus is rapidly growing, as it is simultaneously being performed in two different institutes of computer sciences, in Bucharest and in Iaşi.

3.1. DeReKo

The German Reference Corpus DeReKo (Deutsches Referenzkorpus) has been developed at the IDS since its inception in 1964. With more than 25 billion words (Kupietz and Lüngen, 2014), it is the world's largest collection of German texts. In contrast to other reference or national corpora, DeReKo is not designed to be used as a monolithic corpus. Instead, it adopts a primordial-sample design approach (Kupietz *et al.*, 2010), which invites users to create stratified sub-samples (referred to as virtual corpora or virtual collections), custom-tailored to their respective research questions and basic populations. Such an approach effectively allows for maximization of its size, diversity and applicability for different research questions (Kupietz *et al.*, 2014) and is also fundamental for the definition of different virtual comparable corpora in the DRuKoLA-context. DeReKo provides a broad variety of text types with a quantitative focus on newspaper texts and rapidly growing portion of computer mediated communication (cf. Beißwenger *et al.*, 2015; Margaretha and Lüngen, 2014; Schröck and Lüngen, 2015). DeReKo is endowed with rich metadata (Klosa *et al.*, 2012; Kupietz and Keibel, 2009), multiply annotated on the part-of-speech, dependency and constituency levels (Belica *et al.*, 2011) and sufficiently licensed to be queried and analyzed for non-commercial linguistic purposes (QAO-NC license, see Kupietz and Lüngen, 2014).

3.2. CoRoLa

Currently, CoRoLa contains more than 191 million word forms of written text and about 135 hours of transcribed speech (Tufiş *et al.*, 2016). In its first public version, CoRoLa will contain more than 500 million word forms and more than 300 hours of transcribed speech (approximately 3 million words) and it will be IPR (Intellectual Property Rights) cleared. It aims at being representative for the literary language. The corpus covers the following 35 subdomains: *literature, politics, gossip, film, music, economy, health, linguistics, theatre, painting/drawing, law, sport, education, history, religious studies and theology, medicine, technology, chemistry, entertainment, environment, architecture, engineering, pharmacology, art history, administration, enology, pedagogy, philology, juridical sciences, biology, social, mathematics, social events, philosophy, other*[2]. The domains and sub-domains classification is based on the Wikipedia one. The functional styles considered are*: journalistic, scientific, imaginative, memorialistic, administrative*, *juridic* and *other* (see footnote 2). CoRoLa uses similar realisation conventions as the Romanian Balanced Corpus (ROMBAC)[3] (Ion *et al.*, 2012) containing over 44 million tokens from five domains (*news, medical, legal, biographic* and *fiction*). The creators of CoRoLa pay special attention in obtaining the consent of owners before including their texts in the corpus; thus, protocols of collaboration have been signed with a number of publishing houses, editorial offices, and radio channels.

In line with the current diversification of language and

[2]This is a category for all documents that could not be definitely classified into the named categories.

[3]http://www.meta-net.eu/meta-share

speech information available in modern representative corpora, CoRoLa will include a syntactically annotated subcorpus and an oral component. All textual data is morpholexically processed (tokenized, POS-tagged and lemmatized). The current annotations are provided in-line but, in the future, as different layers of linguistic annotations (noun phrases, dependency parses, name entities, semantic relations, discourse structures etc.) will be provided for the same data, a mixed (in-line and standoff) annotation will be used. The Universal Dependency (UD)[4] compliant treebank (targeted: more than 10.000 hand validated sentences) and the oral component have additional annotations (dependency links, respectively speech segmentation at sentence level, pauses, non-lexical sounds, like breath, cough, laugh, sneeze, and partial explicit marking of the accent).
The metadata annotators (many of which are volunteers) work under the guidance of a detailed Annotation Manual. Started two years before the initiation of DRuKoLA, the work till now devoted in building CoRoLa was technically supported by an online platform (developed at IIT-Iaşi), which includes facilities for cleaning formatting, standardising Romanian diacritics, eliminating hyphenation, visualizing statistics about the quantity of texts accumulated and their subdomains, and filling in metadata. However, many clearing phases are still done manually: separating articles from periodicals in different files, removal of headers, page numbers, figures, tables, text fragments in foreign languages, excerpts from other authors, and annotation of footers and endnotes (decided to be left in the texts).

3.3. Harmonization of DeReKo and CoRoLa

Both CoRoLa and DeReKo metadata comply with CMDI (Component MetaData Infrastructure)[5] and/or TEI-P5[6] standards. For the construction of comparable corpora, however, in addition to mainly syntactical interoperability, also semantic interoperability has to be achieved, for example for the metadata categories that are used for the construction of virtual corpora. The general procedure for the harmonization of data categories and value sets will be to define functions that map the original respective data to more coarse-grained taxonomies. Additional harmonization work will also be required on lower levels, e. g. for the integration of CoRoLa into the KorAP corpus query engine, or for the adoption of the GGS query mechanism developed for CoRoLa as an auxiliary search engine to express constraints that would exploit the multi-layered annotation of DeReKo, both mentioned in the following section. The first DRuKoLA workshop[7] is expected to answer many of these questions.

4. Query and analysis software

The software that will be used for conducting the corpus linguistic research within DRuKoLA and for making the project results available to the community is the corpus query- and analysis platform KorAP that has recently been developed at the IDS (Bański *et al.*, 2013; 2014). KorAP is the designated successor of the corpus search and management system COSMAS that was launched in 1994 and in its second incarnation (COSMAS II), is currently used by 39.000 German linguists. Besides KorAP's more performance oriented features, like horizontal scalability with respect to an unbounded corpus size and any number of annotation layers, two are particularly fundamental for DRuKoLA: 1.) its ability to manage corpora that are physically located at different places, in order to comply with typical license restrictions (cf. Kupietz *et al.*, 2014) and 2.) its ability to dynamically create virtual sub-corpora based on text properties and to manage these virtual corpora in a persistent way, to e. g. allow for reusability and reproducibility. Further features that will be required for the rather mono-linguistic research purposes will be integrated from recent and ongoing developments of the project partners, as for example the interactive overview visualizations of corpus compositions (Perkuhn and Kupietz, forthcoming), or the visualisation of query expressions as graphs, allowed by the GGS mechanism (Simionescu, 2012). GGS (Graphical Grammar Studio) is an open-source platform allowing interactive writing of grammars that annotate sequences of XML elements at any levels and which has been recently augmented with a constraint-based search mechanism (Simionescu, forthcoming). Also functionalities specifically required for the contrastive research tasks will first be inventoried and then developed during the project.

5. Corpus based contrastive case studies

Based on recent or current research interests of the participating linguists on definite DPs in Romanian (Cornilescu and Nicolae, 2011a; 2011b), situational use of demonstratives (Cosma and Engelberg, 2014) or particularities of the Romanian verbal supine form (Cornilescu and Cosma, 2013; 2014) the project is primarily sustained – as part of the harmonization process – in the making and adapting analyzing instruments for Romanian. The testing phase of the developed instruments will then serve data-based linguistic research and will help identify linguistic variation and preferences within selected research topics: modality markers *haben-zu* infinitives in German and their equivalent finite and nonfinite forms (*are de* V-*ut*, *are* $V_{infinitive}$, *are să* $V_{subj.}$) in Romanian, demonstratives in different uses and positions, reinforcement patterns of demonstratives through adverbs as in *dieser hier*, *dieses schöne Auto da*, propositional reference of demonstratives (*das*, *asta*) etc. Therefore possible aspects to be syntactically explored include: i) distributional patterns of *haben-zu* infinitives with *haben* as a raising verb, distribution of the equivalent form variants of Romanian *are de*/*a*/ *să* + V; ii) identifying structural and stylistic factors in the use of one of the three equivalent forms of the *haben-zu* infinitive in Romanian; iii) the use of propositional demonstrative *das* and singular and plural form differentiated abstract demonstratives *asta*/*astea* in Romanian, etc. For the exploration and analysis of distributional semantic and syntagmatic properties we will use collocation profiles (Belica *et al.*, 2010; Belica, 2011) as well as word embeddings (Mikolov *et al.*, 2013; Ling *et al.*, 2015).

[4] http://universaldependencies.github.io/docs/
[5] http://www.clarin.eu/content/component-metadata
[6] http://www.tei-c.org/Guidelines/P5/
[7] The workshop takes place in April this year in Bucharest

6. Conclusions

We have presented in this paper a very young German-Romanian project, intended to harmonize methods and tools for building and exploiting corpora in these two languages. The idea of the project is to apply a long-standing tradition in the creation of corpora to a newly-born one. At one pole of this project there is the experience gained by the IDS Mannheim in the creation of DeReKo, the largest German language corpus. Two years before this project was initiated, the work on the Contemporary Romanian Language Corpus was simultaneously started in Bucharest and Iaşi. The experience gathered in this period (find providers of texts and vocal recordings, agree on the metadata being used, design and build an interactive platform that helps to clean the linguistic data and fill-in metadata, and design an access mechanism) will now have to be harmonised with the already running German machine. Whether one common methodology will be applicable to both corpora, comparable conventions will have to be fixed through an updating process. This will not only make possible extremely interesting contrastive studies over the two languages and will produce a very large comparative bilingual corpus (with interesting possible beneficials for the MT technology), but the lessons learned from this enterprise could be extended at the European level, to prepare the stage for a multilingual unification of corpora, methodologically and technologically, with tremendous beneficial effects in the multilingual language research.

7. References

Altenberg, B. (1999). Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hasselgård and Oksemell (eds.) *Out of Corpora*. Amsterdam: Rodopi, 249-268.

Bański, P., Bingel, J., Diewald, N. , Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C. and Witt, A. (2013). KorAP: the new corpus analysis platform at IDS Mannheim. Presented at the 6th Conference on Language and Technology (LTC-2013), Poznań, Polen, December 2013.

Bański, P., Diewald, N., Hanl, M., Kupietz, M. and Witt, A. (2014). Access Control by Query Rewriting: the Case of KorAP. In: *Proceedings of the 9th conference on the Language Resources and Evaluation Conference* (LREC 2014), European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014. 3817-3822.

Beißwenger, M., Ehrhardt, E., Horbach, A., Lüngen, H., Steffen, D. and Storrer, A. (2015). Adding Value to CMC Corpora: CLARINification and Part-of-speech Annotation of the Dortmund Chat Corpus In: Beißwenger, M. and Zesch, T. (eds.): NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop , September 29, 2015 University of Duisburg-Essen, Campus Essen. S. 12-16 - : German Society for Computational Linguistics & Language Technology (GSCL), 2015.

Belica, C., Keibel, H., Kupietz, M. and Perkuhn, R. (2010). An empiricist's view of the ontology of lexical-semantic relations. In: Storjohann, P. (ed.) *Lexical-Semantic Relations. Theoretical and practical perspectives*. John Benjamins Publishing Company. 115-144.

Belica, C. (2011). Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In: Abel, A., Zanin, R. (eds.): *Korpora in Lehre und Forschung*, S. 155-178. Bozen-Bolzano University Press. Freie Universität Bozen-Bolzano.

Belica, C., Kupietz, M., Witt, A. and Lüngen, H. (2011). The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Konopka, M., Kubczak, J., Mair, C., Šticha, F., Waßner, U. (eds.): *Grammar and Corpora* 2009. Third international conference. Tübingen: Narr. 451-469.

Chesterman, A. (1998). Contrastive Functional Analysis. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Cornilescu A., Nicolae, A. (2011a). Nominal Peripheries and Phase Structure in the Romanian DP. In: *Revue Roumaine de Linguistique* LVI, 1., 35–68.

Cornilescu, A., Nicolae, A. (2011b). On the Syntax of the Romanian Definite Phrases: Changes in the Patterns of Definiteness Checking. In: Sleeman, P., Perridon H. (eds.): *The Noun Phrase in Romance and Germanic. Structure, Variation and Change*. Amsterdam: John Benjamins. 193−222.

Cornilescu, A., Cosma, R. (2013). Restructuring strategies as means of providing increased referentiality for the internal argument of the de-supine clause. In *Bucharest Working Papers in Linguistics* vol. XV.2., 91-121.

Cornilescu, A., Cosma, R. (2014). On the functional structure of the Romanian de-supine. In: Cosma, R., Engelberg, S., Schlotthauer, S., Stănescu, S., Zifonun, G. (eds.): *Komplexe Argumentstrukturen. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen*. Berlin/München/Boston: de Gruyter. [Konvergenz und Divergenz 3]. 283-335.

Cosma, R., Engelberg, S. (2014). Subjektsätze als alternative Argumentrealisierungen im Deutschen und Rumänischen. Eine kontrastive quantitative Korpusstudie zu Psych-Verben. In: Cosma, R., Engelberg, S., Schlotthauer, S., Stănescu, S., Zifonun, G. (eds.): *Komplexe Argumentstrukturen. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen.* Berlin/München/Boston: de Gruyter. 339- 420.

Ion, R., Irimia, E., Ştefănescu, D. and Tufiş, D. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In Calzolari, Nicoletta et al. (eds.). *Proceedings of the 8th LREC*. 339-344.

Johansson, S. (1999). Corpora and contrastive studies. In Pietilä, P. and Salo, O.-P. (eds.): *Multiple Languages – Multiple Perspectives*. AFinLA Yearbook 1999 / No. 57, 116-125.

Kaczmarska, E., Rosen, A. (2013). Między znaczeniem leksykalnym a walencją – próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego. *Studia z Filologii Polskiej i Słowiańskiej,* 48: 103–121. Warszawa.

Klosa, A., Kupietz, M., and Lüngen, H. (2012). Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica 28. Berlin/Boston: de Gruyter, 71–97.

Kupietz, M. and Keibel, H. (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji

(Eds.): Working Papers in Corpus-based Linguistics and Language Education, No. 3. - Tokyo: Tokyo University of Foreign Studies, 53–59.

Kupietz, M., Belica, C., Keibel, H. and Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, N. et al. (eds.): *Proceedings of LREC 2010*. 1848-1854.

Kupietz, M., Lüngen, 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., Lüngen, H., Bański, P. and Belica, C. (2014). Maximizing the Potential of Very Large Corpora. In: Kupietz, M., Biber, H., Lüngen, H., Bański, P., Breiteneder, E., Mörth, K., Witt, A., Takhsha, J. (eds.): *Proceedings of the LREC-2014*-Workshop Challenges in the Management of Large Corpora (CMLC2). Reykjavik: ELRA, 1–6.

Ling, W., Dyer, C., Black, A. and Trancoso, I. (2015). Two/Too Simple Adaptations of word2vec for Syntax Problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, CO: ACL.

Margaretha, E., Lüngen, H. (2014). Building Linguistic Corpora from Wikipedia Articles and Discussions. In: Beißwenger, M., Oostdijk, N., Storrer, A., van den Heuvel, H. (eds.): *Journal for Language Technology and Computational Linguistics* (JLCL) 29 (2). Special Issue on Building and Annotating Corpora of Computer-mediated Communication: Issues and Challenges at the Interface between Computational and Corpus Linguistics. Regensburg: GSCL, 2014, 59-82.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013): Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS* (Advances in Neural Information Processing Systems) 2013, 3111–3119.

Perkuhn, R., Kupietz, M. (forthcoming): Visualisierung als erkenntnisleitendes Instrument. In Bubenhofer, N. and Kupietz, M.: *Proceedings of the Herrenhausen-Symposium on Visual Linguistics* 2014.

Schröck, J., Lüngen, H. (2015): Building and Annotating a Corpus of German-Language Newsgroups In: Beißwenger, M., Zesch, T. (ed.): NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop, September 29, 2015 University of Duisburg-Essen, Campus Essen. German Society for Computational Linguistics & Language Technology (GSCL), 2015, 17–22.

Sharoff, S., Rapp, R., Zweigenbaum, P. and Fung, P. (eds.) (2013). Building and Using Comparable Corpora. Springer.

Simionescu, R. (2012): Romanian Deep Noun Phrase Chunking Using Graphical Grammar Studio. In *Proceedings of the Conference "Linguistic Resources and Instruments for Romanian Language - ConsILR-2011"*, Bucharest, "Alexandru Ioan Cuza" University of Iaşi Editing House, 135–143.

Simionescu, R. (forthcoming): Symbolic Mechanisms for Describing Linguistic Constraints. Ph.D. Thesis, "Alexandru Ioan Cuza" University of Iaşi.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ş. D., Boroş, T., Teodorescu, N. H., Cristea, D., Scutelnicu, A., Bolea, C., Moruz, A. and Pistol, L. (2015): CoRoLa Starts Blooming – An Update on the Reference Corpus of Contemporary Romanian Language. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora* (CMLC-3), 5-10.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ş., D., Boroş, T. (2016): The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portoroz, Slovenia.

Wälchli, B. (2007): Advantages and disadvantages of using parallel texts in typological investigations. In: *Sprachtypologie und Universalienforschung* 60:2. 118-134.