THE DRUKOLA PROJECT

MARC KUPIETZ[1], RUXANDRA COSMA[2], ANDREAS WITT[1]

Abstract. DRuKoLA, the accompanying project in the making of the Corpus of Romanian Language, is a cooperation between German and Romanian computer scientists, corpus linguists and linguists, aiming at linking reference corpora of European languages under one corpus analysis tool able to manage big data. KorAP, the analysis tool developed at the Leibniz Institute for the German Language (Mannheim), is being tailored for the Romanian language in a first attempt to reunite reference corpora under the EuReCo initiative, detailed in this paper. The paper describes the necessary steps of harmonization within KorAP and the corpus of Romanian language and discusses, as one important goal of this project, criteria and ways to build virtual comparable corpora to be used for contrastive linguistic analyses.

Keywords: comparable corpora, corpus analysis tools, reference corpora, research infrastructure.

1. INTRODUCTION

This paper is dedicated to a project funded by the Alexander von Humboldt-Foundation within the period 2016–2018, named *Contrastive linguistics from a language-technological perspective*[3]. The aim of the project is to transcend the known limits of languages in contrast through corpus technology. This combination of scientific disciplines is not a novelty per se: the potential of corpora in translation studies or in learner corpora for foreign language teaching and second language acquisition has been an important research topic in applied linguistics for more than the last two decades (a.o. Baker 1993, Leech 1997, Granger *et al*. 2002, Hansen-Schirra and Teich 2002, Sinclair 2004, Granger 2008). The idea behind this project is nevertheless new, as it focuses on contrastive linguistics as a journey's end by i. networking, ii. technology and iii. technological networking, whereas i. refers to reference corpora of European languages, ii. to the development of a common analysis platform and of comparable corpora, iii. the link between existing language technologies within the European languages included in the

[1] Leibniz Institute for the German Language, Mannheim, kupietz@ids-mannheim.de, witt@ids-mannheim.de

[2] University of Bucharest, ruxandra.cosma@lls.unibuc.ro

[3] *Sprachvergleich korpustechnologisch. Deutsch-Rumänisch* is the original German title of the project funded as a Research Group Linkage Programme by the Alexander von Humboldt-Foundation for the time span 2016-2018, that DruKoLA stands for. The abbreviation of the project title, derived from *Deutsch-Rumänische korpuslinguistische Analyse,* mentions only one of the goals of the project, namely the ability of doing contrastive research by linking and networking technology.

RRL, **LXIV**, *3*, p. 255–263, Bucureşti, 2019

project. This is not only to be seen from the perspective of building a software dedicated to scientific research but as a research project in itself, as it will be shown next.

2. FOUNDATIONS OF DRUKOLA

For one thing, DRuKoLA is the accompanying project in the making of CoRoLa, the Romanian reference corpus described in this volume. For another, it is about the adaptation of the corpus analysis tool KorAP, originally developed for analyzing the German language, to the needs of a second large community of linguists and to the needs of contrastive linguistic studies.

KorAP (Bański *et al.* 2012, 2013; Diewald *et al.* 2016) was started 2011 as a corpus analysis platform at the Leibniz Institute for the German Language (Leibniz-Institut für Deutsche Sprache, IDS) with a view to the future, as the Reference Corpus of the German Language, DeReKo (Deutsches Referenzkorpus), was growing fast, currently to 43 billion words (Kupietz *et al.* 2018a). Its advantage lies in the fact that the IDS has had a language data and corpus tradition over the last 50 years (Teubert and Belica 2014). On the development phases of CoRoLa we refer to Tufiş *et al.* and Gîfu *et al.* (both in this volume).

Due to the elaboration, progress and research phase at the beginning of the project of the Romanian reference corpus and of the analysis tool KorAP, this was a perfect time to embrace the idea of linking, under one instrument, two reference corpora situated in home institutes in Germany and Romania (Bucharest, Jassy). For CoRoLa, KorAP is the main analysis tool, yet, at the same time, one of three options for querying the reference corpus.

But DRuKoLA is to be understood in breach of the limits of a two-sided contrastive project. As a basis for a two-sided contrastive research project, it offers one technological instrument for two languages, with respect to interrogation and construction of comparable data in terms of corpus architecture and management. However, DRuKoLA is from more than one perspective a road opener. The pioneering work on harmonizing and adapting the scientific software developed for the analysis of the German language to a Romance language, in support of the idea of an abstract template in language structure, is one thing. The novelty of this work consists in training to open to other reference corpora of other languages, as well. This would be the idea of EuReCo, a network of European reference corpora, locally situated in their home countries, able to be analyzed under one analysis tool. Therefore, DRuKoLA is about change in contrastive corpus-linguistics and about technological impetus.

DRuKoLA is intrinsically a collaborative project. As initiators and developers of the language analysis platform, standing behind the idea of EuReCo, there is the Leibniz Institute for the German Language (IDS), represented by corpus and computational linguists Marc Kupietz, Nils Diewald, Eliza Margaretha Illig and Andreas Witt. On the other side, in building, expanding and providing the reference corpus of Romanian, CoRoLa, there are two institutes of the Romanian Academy engaged in – the *Nicolae Drăgănescu* Research Institute for Artificial Intelligence (RACAI), with team members in DRuKoLA Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia and Ştefan Dumitrescu as computer scientists and linguists, and the Institute of Computer Science in Jassy (IIT), represented in DruKoLA by a team of three computer scientists – Dan Cristea, Andrei Scutelnicu, Alex Moruz. Linguists from the University of Bucharest – Ruxandra Cosma,

Alexandra Cornilescu, Maura Cotfas, Vlad Cucu Oancea (from the German and the English Department), tested the platform KorAP on Romanian and explored language data provided by CoRoLa at different project stages (see the papers in this volume).

3. THE EURECO VISION

The past 20 years have seen an emergence of national, reference and other large monolingual corpora of numerous European languages (cf. Kupietz *et al.* 2017). Most of them have been or are being built in projects of limited duration, but typically located at institutions that are at least to some degree responsible for curating data and for making it available to the respective scientific communities also after the building phase. The main idea of the EuReCo initiative (Kupietz *et al.* 2017) is that such institutions should join forces for the development of research software and the missing parts of research infrastructure to allow for a unified use of the existing corpora in a sustainable and economically feasible way. EuReCo wants to achieve this by joining the existing corpus resources just virtually, keeping them physically at their host institutions. The expected positive effects of this approach are:

- IPR and licensing issues can be avoided because the typically legally relevant full texts can physically stay at their original host institutions.
- The virtually joint corpora can automatically benefit from further developments, maintenance and expansion of the virtually integrated individual corpora.
- Development and maintenance costs for research software can be distributed.
- Corpora will be reusable in more application areas.
- Virtual *comparable* corpora can be defined for different language pairs based on the integrated national and reference corpora, so that the strong demand for multilingual corpora of high linguistic quality (with respect to size, dispersion, the richness of metadata and linguistic annotations) and high comparability can be alleviated (see Kupietz *et al.* 2018b).
- Corpora of different languages can be used and analyzed in a unified fashion.
- The joint use and development of a corpus analysis platform might counteract methodical isolation of the individual philologies.

3.1. Infrastructural Aspects

In addition, EuReCo can not only build upon and reuse existing corpus resources but also on the base infrastructures already built by the CLARIN initiative. After the experience with the CLARIN project it has to be noted, however, that also the EuReCo infrastructure is unlikely to address the demands of all desired research applications of the integrated corpora. The experience with infrastructure research for linguistic data has shown that against the backdrop of the current copyright and data protection legislation, within the digital humanities and even in linguistics alone the applications are too diverse and heterogeneous to be fully covered by a common research infrastructure (Kupietz *et al.* 2018c). On the other hand, CLARIN and DRuKoLA have shown that a large part of these applications shares many common sub-tasks that can indeed be supported by more or less

general infrastructures. While CLARIN focuses on a base layer of data and metadata exchange standards and basic services, such as authorization and authentication as well as registration of persistent identifiers, EuReCo builds upon this base layer a more specialized column of functionalities that can be useful for institutions to make large corpora available to (corpus) linguists. As indicated above, typically, more of such columns will be needed to provide all desired functionalities, because a jack of all trades research infrastructure, reaching vertically up to end-users, is, if not impossible, not feasible and not maintainable in practice.

3.2 Further EuReCo and related projects

As the second project in the EuReCo context, DeutUng (*Comparing German and Hungarian: corpus-technological, functional-semantic and language didactic perspectives*) has started in 2017 with the integration of the Hungarian National Corpus HNC (Váradi 2002; Oravecz *et al.* 2014). DeutUng is a cooperation project between the IDS and the University of Szeged (Hungary) with the Research Institute for Linguistics at the Hungarian Academy of Sciences as associated partner, also funded by the Alexander von Humboldt-Foundation as a Research Group Linkage Programme. Besides the EuReCo-integration of the HNC, DeutUng's goals also comprise the construction of an error-annotated German-Hungarian learner corpus DULKO (Hirschmann and Nolda 2019), as well as a comparison of selected grammatical phenomena.

After the integration of the HNC, EuReCo will be ready to invite more national and reference corpora. Concrete plans already exist for the integration of the National Corpus of Polish NKJP (*Narodowy Korpus Języka Polskiego*) (Przepiórkowski *et al.* 2012) in collaboration with the Polish Academy, that was also one of the initial partners of the EuReCo initiative in 2013.

4. IMPLEMENTATION OF THE DRUKOLA PROJECT

In order to provide a dynamic design of smaller sub-corpora to be compared to each other, harmonization and iteration were the key-concepts followed in the creation of distinct comparable corpora and in the linking-process. These principles are described in the next sections of this chapter. Harmonization of corpus resources would mainly mean i. harmonization of annotation principles, ii. mapping of metadata, iii. feed strata which hadn't been sufficiently fostered before, iv. add further morphological and syntactical annotation layers, if needed. Based on a comparable architecture of CoRoLa and DeReKo and on the TEI-Standard followed, only basic harmonization work was required, like linking the text-taxonomy of CoRoLa to that of DeReKo, with no further need of expansion.

4.1. General approach to "good" comparable corpora

The question of how to define and construct comparable corpora has been one of the most important and most fundamental tasks within the project. To tackle it, our plan was to use an iterative bootstrapping approach to stepwise develop eventually multiple scores of comparability and at the same time to optimize them in different tracks, resulting in

different comparable corpora that are gradually improved with each iteration. Such an iterative approach to improving the representativeness of a corpus with respect to a targeted language domain and research question is in general advisable to avoid artifacts caused by skewed sample compositions (Kupietz 2015). With comparable corpora, however, it even seems necessary to take such an approach because the additional corpus in a second language and the additional requirement of comparability add two factors to the risk of artifacts caused by skewed sample compositions. Moreover, the question of what a comparable corpus means is a very fundamental research question that, in the context of linguistic research, will have no single and no final answer.

Roughly paraphrased, a good comparable corpus will be one that allows the inference from quantitative corpus findings to an answer of a contrastive linguistic research question, with the condition that the answer is not an effect of a skewed L1-corpus, L2-corpus or comparability-relation. As none of these conditions can be formalized, they can only be approached empirically for individual cases by varying the comparability and sampling criteria. Thus, constructing comparable corpora is here regarded as an extension of approaching individual representativeness in the construction of monolingual corpora by iterative, stratified sub-sampling, described in detail in Kupietz (2015):

1. start with a good mapping of metadata properties
2. define a comparable corpus pair
3. perform comparative case studies
4. refine mapping, if findings (or effect sizes) could be artifacts of skewed corpus compositions and start over with 2

With the possibility of defining and refining comparability criteria and thereby comparable corpus pairs dynamically, also the stability of quantitative findings with regard to differently defined comparable corpora can be evaluated. It has to be noted, however, that the flexibility of different comparable corpus definitions is limited by the size and stratification of the underlying monolingual corpora and that additional comparability criteria will typically reduce the size of the resulting comparable corpus pairs so that the approach cannot avoid a tradeoff between comparability and corpus size.

4.2. Practical constraints to comparability criteria

Within the DRuKoLA project we were not able to fully put our theoretical approach into practice allowing also end-users to play around dynamically with comparability criteria. The main reason was that this would have required a sampling function in KorAPs corpus composition editor to allow for downsampling the strata defined by the comparability criteria to equal sizes. As such a function is part of a basic KorAP module for parallel search that has not been fully implemented (see Diewald and Margaretha 2016: 87), we decided to first apply our approach outside of KorAP and to start with providing only a small number of fixed comparable virtual corpora to KorAP users.

The comparable metadata categories that were available in both, DeReKo and CoRoLa, were publication date, text type and topic domain. We decided to start with topic domain as the first comparability criterion, as for it we had a very high resolution as well as a good distribution of texts to the categories. Both corpora provide a two-level domain classification for every text, however with different and incongruent categories on both levels. While DeReKo's taxonomy is based on the one of the Open Directory Project

(dmoz) (Weiß 2005, Klosa *et al.* 2012), CoRoLa's taxonomy is based on the Universal Decimal Classification (UDC) and the Wikipedia top-level domains (see Gîfu *et al.*, this volume). For this reason, we had to define a mapping between the two taxonomies. Our initial plan was to define a common coarse taxonomy with mappings for both corpora, but it turned out that the most straightforward approach was rather to map CoRoLa'stop- and sub-domains to DeReKo's top- and sub-domains, only. To be able to improve the mapping, we plan to provide UDC and Wikipedia domains for DeReKo, in the future. With the current approach, however, we have already achieved a satisfactory mapping for 99% of the labelled texts in CoRoLa at the top level and for 90% of the labelled texts at the subdomain level.

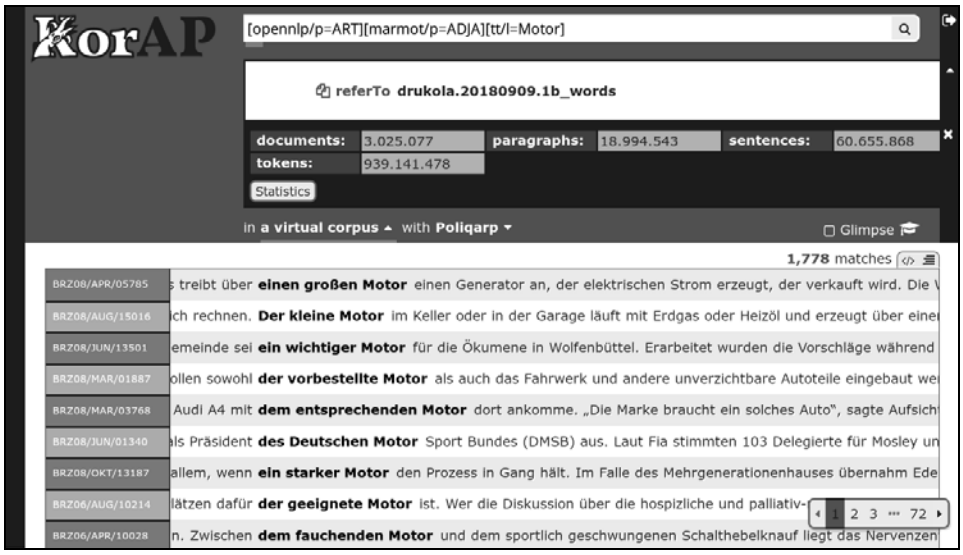

Fig. 1. Referencing to a first persistent virtual comparable German-Romanian corpus in KorAP.

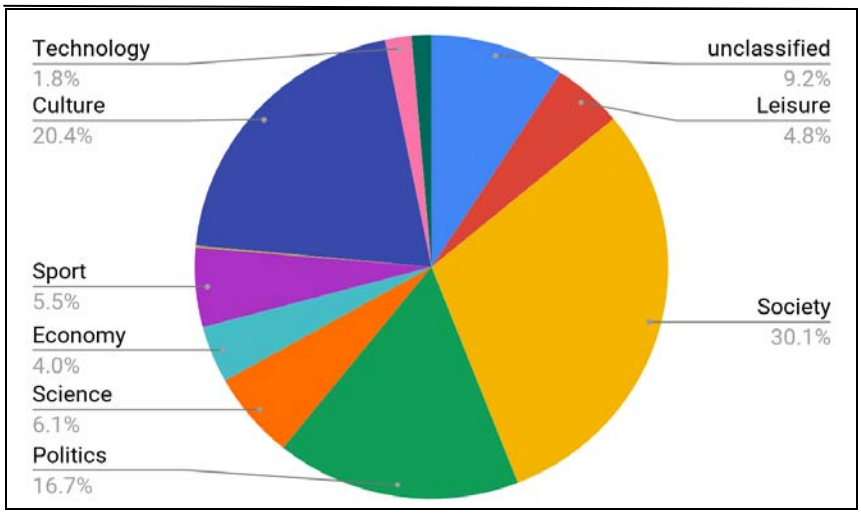

Fig. 2. Distribution of words in the first comparable corpus to DeReKo's top-level topic domains.

4.3. The first comparable German-Romanian corpora

Based on the domain mapping we have drawn a stratified random sample from DeReKo-2018-I so that for each mapped domain the number of texts (and approximately also the number of tokens) from DeReKo was equal to the one from CoRoLa. This virtual DeReKo sub-corpus is stored as a persistent virtual corpus (VC) in KorAP and can be referenced (optionally as part of a more complex VC) to restrict search and analysis to all documents in the comparable corpus.[4] The German part of the German-Romanian comparable corpus consists of more than 3 million documents, comprising 940 million word tokens (see Figure 1) with a topic domain distribution, according to the DeReKo taxonomy, shown in Figure 2.

The next generation of comparable German-Romanian corpora will also take the temporal distribution into account. First experiments show that this will be possible without having to reduce the resulting corpus size due to the additional constraints. Subsequent versions will then also take into account the results of quantitative comparative experiments and possible hints to skewed corpus compositions and thus realize the full comparability approach sketched in section 4.1, above. This last step, however, also depends on the finalization of KorAP's parallelization component which also covers quantitative and aggregation functions. We expect that building comparable corpora from the perspective of specific needs and research objectives under DeReKo and CoRoLa dynamically should be possible by the beginning of 2020.

5. CONCLUSIONS AND OUTLOOK

In disciplines that deal with large amounts of data, it is probably a truism that one person's research is another person's tool or infrastructure or data. But even without taking different perspectives, the boundaries between infrastructure, data, qualitative interpretation and disciplinary research are becoming more and more blurred with the increasing demands on methodologically sound interpretation of the growing amount of empirical data. The interdependencies between components of data grounded research are too strong to be dealt with in isolation from each other. This means that not only quantitative and qualitative linguistic research requires corpus data and tools, but also that the decisions made for the construction of the corpus can have a decisive impact on the quantitative and on the qualitative results, which is especially the case with comparable corpora (see also Cosma and Kupietz 2018: 205). Even more important for DRuKoLA and EuReCo was and is, however, its infrastructural component, without which such an endeavour would only be possible at considerable expense and, accordingly, without the prospect of sustainability. One of our conclusions from the project is that a successful further development of the digital humanities requires not only a closer integration of qualitative and quantitative methods and tools, but also a stronger integration of infrastructural solutions to technical, legal, economical and collaborative challenges. While in the pre-digital ages, libraries were capable of guaranteeing a constant, gradual growth of knowledge and data, nowadays sustainable research infrastructures play a major part in that role. This does not only

[4] https://korap.ids-mannheim.de/?q=Test&collection=referTo+"drukola.20180909.1b_words".

concern the mere preservation of knowledge, or accordingly the long-term preservation of data and tools, but also the creation of platforms, in a technical and in a broader metaphorical sense, from which more advanced research objectives can be reached.

The DRuKoLA project has built the first parts of EuReCo. We hope that EuReCo will grow and serve as a sustainable basis for more researchers to reach out for more advanced objectives in general corpus linguistics and in contrastive studies.

REFERENCES

Baker, M., 1993, "Corpus linguistics and translation studies. Implications and applications", in: M. Baker, G. Francis, E. Tognini-Bonelli (eds), *Text and technology*. Philadelphia / Amsterdam, John Benjamins, 232–252.

Bański, P., P. M. Fischer, Frick, E. Ketzan, M. Kupietz, C. Schnober, O. Schonefeld, A. Witt, 2012, "The New IDS Corpus Analysis Platform: Challenges and Prospects." in: N. Calzolari, K. Choukri, T. Declerck, M. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), Istanbul, Turkey, May 2012. European Language Resources Association (ELRA), 2905–2911.

Bański, P., J. Bingel, N. Diewald, E. Frick, M. Hanl, M. Kupietz, P. Pęzik, C. Schnober, A. Witt, 2013, "KorAP: the new corpus analysis platform at IDS Mannheim", in: Z. Vetulani, H. Uszkoreit (eds), *Human Language Technologies as a Challenge for Computer Science and Linguistics*, *Proceedings of the 6th Language and Technology Conference*, Poznań, Fundacja Uniwersytetuim. A., 586–587.

Barbu Mititelu, V., D. Tufiș, E. Irimia, 2018, "The Reference Corpus of the Contemporary Romanian Language (CoRoLa)", in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds), *Proceedings* of the *Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki / Paris, ELRA, 1178–1185.

Cosma, R., M. Kupietz, 2018, "Von Schienen, Zügen und linguistischen Fragestellungen", in: H. Lobin, R. Schneider, A. Witt (eds), *Digitale Infrastrukturen für die germanistische Forschung* (= Germanistische Sprachwissenschaft um 2020, vol. 6), Berlin / Boston, de Gruyter, 199–218.

Diewald, N., E. Margaretha, 2016, "Krill: KorAP search and analysis engine", in: M. Kupietz, A. Geyken (eds), *Corpus Linguistic Software Tools,* Journal for language technology and computational linguistics (JLCL) 31 (1), Berlin, GSCL, 73–90.

Diewald, N., M. Hanl, E. Margaretha, J. Bingel, M. Kupietz, P. Bański, A. Witt, 2016, "KorAP Architecture – Diving in the Deep Sea of Corpus Data", in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia. Paris, European Language Resources Association (ELRA), 3586–3591.

Granger, S., J. Hung, S. Petch-Tyson, 2002, *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*, Amsterdam, Benjamins.

Granger, S., 2008, "Learner corpora", in: A. Lüdeling, M. Kytö (eds), *Corpus Linguistics, An International Handbook*, Volume 1, Berlin & New York, Walter de Gruyter, 259-275.

Gîfu, D., A. Moruz, C. Bolea, A. Bibiri, M. Mitrofan, 2019, "The Methodology of Building CoRoLa", in this volume.

Hansen-Schirra, S., E. Teich, "Corpora in human translation", in: *Corpus Linguistics. An International Handbook*, Vol. 1, Berlin, de Gruyter, 1159–1175.

Hirschmann, H., A. Nolda, 2019, "Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus", in: L. Eichinger, A. Plewnia (eds), *Neues vom heutigen Deutsch, Empirisch – methodisch – theoretisch*, Institut für Deutsche Sprache, Jahrbuch 2018, Berlin, de Gruyter, 339–342.

Klosa, A., M. Kupietz, H. Lüngen, 2012, "Zum Nutzen von Korpusauszeichnungen für die Lexikographie", *Lexicographica* 28, Berlin / Boston, de Gruyter, 71–97.

Kupietz, M., 2015, "Constructing a Corpus", in: P. Durkin (ed.), *The Oxford Handbook of Lexicography* (= Oxford handbooks in linguistics), Oxford, Oxford University Press, 62–75.

Kupietz, M., A. Witt, P. Bański,, D. Tufiş, D. Cristea, T. Váradi, 2017, "EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research", in: P. Bański, M. Kupietz, H. Lüngen, P. Rayson, H. Biber, E. Breiteneder, S. Clematide, J. Mariani, M. Stevenson, T. Sick (eds), *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing*, Mannheim, IDS, 15–19.

Kupietz, M., H. Lüngen, P. Kamocki, A. Witt, 2018a, "The German Reference Corpus DeReKo: New Developments – New Opportunities", in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, European Language Resources Association (ELRA), 4353–4360.

Kupietz, M., R. Cosma, D. Cristea, N. Diewald, B. Trawiński, D. Tufiş, T. Váradi, A. Wöllstein, 2018b, "Recent developments in the European Reference Corpus (EuReCo)", in: S. Granger, M.-A. Lefer, L. Aguiar de Souza Penha Marion (eds), *Using Corpora in Contrastive and Translation Studies Conference*, (5th edition), Book of Abstract, Louvain-la-Neuve, CECL, 101–103.

Kupietz, M., N. Diewald, P. Fankhauser, 2018c, "How to Get the Computation Near the Data: Improving data accessibility to, and reusability of analysis functions in corpus query platforms", in: P. Bański, M. Kupietz, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, A. Witt (eds), *Corpora (CMLC-6)*. 07 May 2018 – Miyazaki, Japan. Paris, European language resources association (ELRA), 20–25.

Leech, G., 1997, "Teaching and language corpora: A convergence", in: A. Wichmann, S. Fligelstone, T. McEnery, G. Knowles (eds), *Teaching and language corpora*, London, Longman, 1–23.

Oravecz, Cs., T. Váradi, B. Sass, 2014, "The Hungarian Gigaword Corpus", in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of* Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik / Paris, ELRA, 1719–1723.

Przepiórkowski, A., M. Bańko, R. L. Górski, B. Lewandowska-Tomaszczyk, M. Łaziński, P. Pęzik, 2011, "National Corpus of Polish", *Proceedings of the 5th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, 259–263.

Sinclair, J. (ed.), 2004, *How to Use Corpora in Language Teaching*, Amsterdam, Benjamins.

Teubert, W., C. Belica, 2014, "Von der linguistischen Datenverarbeitung am IDS zur 'Mannheimer Schule der Korpuslinguistik'", in: Institut für Deutsche Sprache (eds), *Ansichten und Einsichten, 50 Jahre Institut für Deutsche Sprache.* Mannheim, Institut für Deutsche Sprache, 298–319.

Tufiş, D., V. Barbu Mititelu, E. Irimia, V. Păiş, R. Ion, N. Diewald, M. Mitrofan, M. Onofrei, 2019, "Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus for Contemporary Romanian", in this volume.

Váradi, T., 2002, "The Hungarian National Corpus", *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC 2002), Las Palmas / Paris: ELRA, 385–389.

Weiß, C., 2005, "Die thematische Erschließung von Sprachkorpora", *OPAL – Online publizierte Arbeiten zur Linguistik*, 1, Mannheim, Institut für Deutsche Sprache.