Erschienen in: Granger, Sylviane/Lefer, Marie-Aude/Aguiar de Souza Penha Marion, Laura (Hrsg.): Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition). – Louvain-la Neuve: CECL Papers 1, 2018. S. 101-103.

Recent developments in the European Reference Corpus (EuReCo)

Marc Kupietz[1], Ruxandra Cosma[2], Dan Cristea[3,4], Nils Diewald[1], Beata Trawiński[1], Dan Tufiş[5], Tamás Váradi[6], Angelika Wöllstein[1]
Institut für Deutsche Sprache[1], University of Bucharest[2],
Romanian Academy Iaşi[3], "Alexandru Ioan Cuza" University of Iaşi[4],
Institute for Artificial Intelligence Mihai Drăgănescu[5], Hungarian Academy of Sciences[6]
kupietz@ids-mannheim.de, ruxandra.cosma@lls.unibuc.ro, dcristea@info.uaic.ro,
diewald@ids-mannheim.de, trawinski@ids-mannheim.de, tufis@racai.ro,
varadi.tamas@nytud.mta.hu, woellstein@ids-mannheim.de

Introduction

The past 20 years have seen an emergence of national, reference and other large 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 based 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 idea of EuReCo (Kupietz et al. 2017) is that such institutions should join forces to develop techniques that allow for a unified view on the existing corpora and to use them as a base for comparable corpora. The common infrastructure will include the following main features: metadata shall use attributes and values that are mappable among all pairs of component corpora, annotation conventions shall be harmonised (as much as linguistic idiosyncrasies permit) to the point that comparable queries on different languages shall be possible and shall produce comparable results, textual content of the component corpora can remain at their hosting institution and be locally extended (quantitatively), updated (w.r.t textual data) or upgraded (w.r.t. metadata and annotation), any component can be accessed from anywhere through a Portal entry point, mixed screens combining more comparable searched for material can be activated dynamically, tools doing statistical counts can be invoked by users on any language component and on any combinations of them, tools used in statistical counts and comparisons shall be able to combine flows of data contributed by all local linguistic data hosts, etc. The expected advantages of this approach are that no comparable corpora would have to be built from scratch, all existing corpora can remain at their hosting institutions – avoiding IPR and licensing issues – and the base for the selection of comparable pairs of sub-corpora could directly benefit from the expansion of the individual initial corpora. The downside of this approach, however, compared e.g. to the similar approach of the ICC (Kirk & Čermáková 2017) is of course that the stratification and composition of possible comparable corpus pairs cannot be designed in advance, but rather depends on the strata manifested by metadata in the source corpora, their respective sizes and the translatability of these stratifications or rather metadata taxonomies between individual corpus pairs.

Previous and current work

EuReCo is currently based on the following corpora:

- The German Reference Corpus DeReKo (Deutsches Referenzkorpus), with more than 42 billion words (Kupietz & Lüngen 2014; Kupietz et al. to appear), the largest linguistically motivated collection of German texts, featuring a so-called primordial-sample design, which is also fundamental for the definition of different virtual comparable corpora in the EuReCo context.
- The Reference Corpus of Contemporary Romanian Language CoRoLa, containing almost one billion words, which was publicly launched in December 2017 and can be queried via different interfaces, including KorAP.[1]
- The Hungarian National Corpus HNC, that has recently been substantially upgraded and extended to gigaword size (Váradi 2002; Oravecz et al. 2014).

KorAP

[1] http://corola.racai.ro/

The current technical basis for EuReCo is the corpus query and analysis platform KorAP[2] that has recently been developed at the IDS (Bański et al. 2013; Diewald et al. 2016). KorAP is the designated successor of the corpus search and management system COSMAS,[3] which is currently used by 40,000 researchers working on the German language. The features of KorAP that are essential for EuReCo are particularly (i) 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 (ii) its ability to dynamically create virtual sub-corpora based on text properties and to manage these virtual corpora in a persistent way, for example allow for reusability and reproducibility.

DRuKoLA: The first EuReCo blueprint

Parts of the EuReCo vision have already been implemented in the DRuKoLA-project,[4] which is centered around DeReKo and CoRoLa (Cosma et al. 2016). One of its main objectives is to provide a common platform for constructing various kinds of comparable corpora based on text properties and to analyse them for contrastive linguistic purposes.

The present state of the art of DRuKoLA relevant to EuReCo is that CoRoLa can be accessed publicly via KorAP and that a first virtual comparable corpus is defined. This first definition is based solely on a mapping from CoRoLa's two-level topic domain taxonomy to DeReKo's topic domain taxonomy (also two-levelled, see Klosa et al. 2012: 88). In order to be able to map a sufficiently substantial portion from the smaller corpus CoRoLa, it was necessary to map from top-level domains to sub-level domains and vice versa.

DeutUng

As a second EuReCo pilot project, DeutUng[5] has recently started to integrate the Hungarian National Corpus (HNC) into EuReCo. With respect to the establishment of an infrastructure and research methodology for comparable corpora, DeutUng is similar to DRuKoLA.[6]

Relevance for Cross-Linguistic Research

Cross-linguistic research needs multilingual data. So far, parallel / translational resources have played a major part, both in contrastive linguistics and language typology as well as in translational studies and foreign language education (cf. James 1980; Chesterman 1998; Granger et al. 2003; Granger 2010; Johansson 2007; Cysouw & Wälchli 2007). While the usefulness of parallel resources for cross-linguistic research is obvious (as they provide data that convey the same meaning and can thus serve as a basis for establishing equivalence between entities across different languages), they show a number of undesirable effects. The problems include particularly the so-called *source language shining through* (Teich 2003), and other specifics of translated texts, such as *over-normalization, simplification,* etc. No such problems arise in comparable data. In this respect, EuReCo, which is based on the existing national or reference corpora, provides a unique linguistic resource offering new perspectives for fine grained contrastive research on authentic cross-linguistic data, applications in translation studies and foreign language teaching and learning.

References

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

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

[2] https://korap.ids-mannheim.de/

[3] http://cosmas2.ids-mannheim.de

[4] DRuKoLA (2016-2019) is funded by the Alexander von Humboldt-Foundation, as a Research Group Linkage Programme. The acronym combines central goals of the project: corpus development and contrastive linguistic analysis (*Sprachvergleich korpustechnologisch. Deutsch-Rumänisch*).

[5] DeutUng (2017-2020) is a cooperation project between IDS Mannheim and the University of Szeged with the Research Institute for Linguistics at the Hungarian Academy of Sciences as associated partner. It is also funded by the Alexander von Humboldt-Foundation as a Research Group Linkage Programme.

[6] With respect to linguistic application, however, DeutUng has as an additional focus on second language acquisition.

Cosma, R., Cristea, D., Kupietz, M., Tufiş, D. & Witt, A. (2016). DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora. In P. Bański, M. Kupietz, H. Lüngen, A. Witt, A. Barbaresi, H. Biber, E. Breiteneder & S. Clematide (eds). *4th Workshop on Challenges in the Management of Large Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), May 23-28, Portorož, Solevia. Paris: European Language Resources Association (ELRA), 28-32.

Cysouw M. & Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung* 60(2), 95-99.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P. & Witt, A. (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), May 23-28, Portorož, Solevia. Paris: European Language Resources Association (ELRA), 3586-3591.

Granger, S. (2010). Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University*, 2, 14-21.

Granger, S., Lerot, J. & Petch-Tyson, S. (eds). (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Amsterdam & Atlanta: Rodopi.

James, C. (1980). *Contrastive Analysis*. London: Longman.

Johansson, S. (2007). *Seeing through multilingual corpora. On the use of corpora in contrastive studies*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Kirk, J. & Čermáková, A. (2017). From ICE to ICC: The new International Comparable Corpus. 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: Institut für Deutsche Sprache, 7-12.

Klosa, A., Kupietz, M. & Lüngen, H. (2012). Zum Nutzen von Korpusauszeichnungen für die Lexikographie. *Lexicographica* 28, 71-97.

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds). *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC 2010), May 17-23, Valletta, Malta, 1848-1854.

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

Kupietz, M., Lüngen, H., Kamocki, P. and Witt, A. (to appear in 2018). The German Reference Corpus DeReKo: New developments – new opportunities. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), 7-12 May 2018, Miyazaki, Japan.

Kupietz, M., Witt, A., Bański, P., Tufiş, D., Cristea, D. & Váradi, T. (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: Institut für Deutsche Sprache, 15-19.

Oravecz, Cs., Váradi, T. & Sass, B. (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 on the Ninth International Conference in Language Resources and Evaluation* (LREC 2014), May 26-31, Reykjavik, Iceland. Paris: European Language Resources Association (ELRA), 1719-1723.

Teich, E. (2003). *Cross-linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

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 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), May 23-28, Portorož, Solevia. Paris: European Language Resources Association (ELRA), 2516-2521.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ş. D., Boroş, T., Teodorescu, N. H., Cristea, D., Scutelnicu, A., Bolea, C., Moruz, A. & Pistol, L. (2015). CoRoLa Starts Blooming – An Update on the Reference Corpus of Contemporary Romanian Language. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen & A. Witt (eds). *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora* (CMLC-3). Mannheim: Institut für Deutsche Sprache, 5-10.

Váradi, T. (2002). The Hungarian National Corpus. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas & Paris: European Language Resources Association (ELRA): 385-389.