Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 15-19

EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research

Marc Kupietz[1], Andreas Witt [1,2,3], Piotr Bański[1], Dan Tufiş[4], Dan Cristea[5,6], Tamás Váradi[7]

[1] Institut für Deutsche Sprache, Mannheim
[2] University of Cologne, Faculty of Arts and Humanities
[3] Heidelberg University, Department of Computational Linguistics
[4] Institute for Artificial Intelligence Mihai Drăgănescu, Bucharest
[5] Romanian Academy, Institute for Computer Science - Iaşi
[6] "Alexandru Ioan Cuza" University of Iaşi, Department of Computer Science
[7] Research Institute for Linguistics, Hungarian Academy of Sciences
(kupietz|witt|banski)@ids-mannheim.de, tufis@racai.ro, dcristea@info.uaic.ro, varadi.tamas@nytud.mta.hu

Abstract

In this paper we discuss the opportunities, prerequisites, possible applications and implications of a virtually joint corpus based on existing national, reference or other large corpora and their host institutions.

1 Introduction

The past 20 years have seen an emergence of national, reference and other large corpora of numerous European languages (Aston & Burnard, 1998; Váradi, 2002; CNC, 2005; Geyken, 2007; Baroni et al., 2009; Davies, 2010; Kupietz et al., 2010; Przepiórkowski et al., 2010; Oravecz et al., 2014; Tufiş et al., 2016). 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, which has been around in the CMLC workshop series since 2012 (see Bański et al., 2012), is that such institutions, rather than continuing as "research islands", should join forces and experiment whether a well-designed technology could allow a unifying view on building and exploitation of a multilingual collection of comparable corpora, a goal motivated by the rapidly changing and growing variety of needs of the linguistic and related user communities.

We present in this paper such a joint project, called EuReCo, briefly showing its aims, the technology behind and the language resources involved.

2 Aims

2.1 Comparable corpora

One of the aforementioned growing needs is the need for comparable corpora in order to facilitate contrastive and generally cross-linguistic research beyond the possibilities provided by parallel corpora, which are very much limited for linguistic applications by unavoidable translation biases. This application area is also the initial and currently the main focus of EuReCo. It appears that joining forces in this area is a particularly promising prospect: given that several national and reference corpora are built and maintained anyway and independently, with methodologies and techniques developed for joining them virtually, where each national centre is still responsible for its language and each corpus still physically located at its centre, it should be much more economical, scalable and sustainable to build a single virtual comparable corpus linking these existing resources than to create the comparable corpora from scratch, possibly even at more than one centre.

2.2 Further aims

In the meantime, however, the envisioned EuReCo has acquired a broader range of potential applications: if the organisational and technical prerequisites for such an infrastructure prove feasible, it would be wise to identify – as early as possible – further functionalities that are currently required or envisioned by the collaboration partners, such as, for example: the ability to manage very large corpora, statistical analysis – ideally dynamically offered to the user, or support for querying different kinds of linguistic annotations.

The general goal of the EuReCo initiative is to bring together existing European corpus initiatives,

specifically in those areas where synergy effects can be expected with high certainty and in a very much target-oriented fashion, towards goals that the collaboration partners would like to achieve, but are unlikely to achieve alone in a sufficiently effective and sustainable way.

Apart from these rather economical aspects, EuReCo also expects benefits from bringing closer together research communities that are currently centered around philologies and their sub-disciplines.

2.3 Relation to CLARIN

The EuReCo objectives are much narrower and oriented towards target applications than those of the European Language Resource Infrastructure Project CLARIN, which "*makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access.*" (see www.clarin.eu). In contrast to CLARIN, which has been particularly strong at providing horizontal base layers of infrastructure, standards and best practices, EuReCo will typically aim at vertical columns ending directly at end-user applications.

3 Foundations

Despite the differing scope and objectives, EuReCo will necessarily be tightly integrated into CLARIN. In addition, its roots lie in a number of experiences gathered by its collaboration partners and their respective histories of providing corpora and tools for using them, in large part within CLARIN:

- contemporary corpora are always tied to their hosting organizations by license contracts and other legal restrictions (Kupietz *et al.*, 2014),
- the way linguists use corpus data is itself subject to rapidly developing research,
- the exact requirements of corpus search and analysis tools for different corpora differ with research traditions and target communities,
- there will be no single tool that satisfies all user needs,
- unification is often necessary to keep costs manageable and to allow for re-usability, but one has to be very careful to keep the results usable and useful.

Based on these insights, the EuReCo strategy can be characterized by the following key properties:

- the aims of the collaboration have to be carefully picked and outlined in order to guarantee that:
  - the product of the collaboration is actually useful for the collaboration partners and their research communities,
  - the overhead of the collaboration does not outweigh its synergy effects,
- commonly developed and used tools must acknowledge the fact that the corpus data itself may not leave its hosting organizations,
- the collaboratively developed tools will usually not replace, but only complement those already existing.

4 Previous and current work

4.1 KorAP

The current main technical basis for EuReCo is the corpus query and analysis platform KorAP that has recently been developed at the IDS (Bański *et al.*, 2013; 2014, Diewald *et al.*, 2016). KorAP is the designated successor of the corpus search and management system COSMAS, which was launched in 1994 and in its second incarnation (COSMAS II), is still currently used by 39.000 researchers working on the German language. Besides KorAP's more performance-oriented features, such as horizontal scalability with respect to an unbounded corpus size and any number of annotation layers, two are particularly fundamental for EuReCo: (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, to e. g. allow for reusability and reproducibility. In addition, using a micro-service-like architecture, KorAP has been specifically designed for collaborative development and particularly collaborative extensibility up to the end-user. Extensibility is also KorAP's main approach to Jim Gray's famous postulate "*put the computation near the data",* which is essential not only to cope with big data, but also to cope with intellectual property rights (IPR) restrictions.

4.2 CoRoLa

CoRoLa is a priority project of the Romanian Academy, carried on by the Institute of Artificial Intelligence "Mihai Drăgănescu" in Bucharest and

the Institute for Computer Science in Iaşi, both affiliated with Romanian Academy. When finalised (end of 2017), CoRoLa will be the largest corpus of Romanian contemporary language, including both written and spoken data. The distinctive aspect of the CoRoLa project (Tufiş *et al.*, 2016) is that all the data included into the reference corpus have cleared IPR, based on bilateral agreements between the developing institutions and the data providers. The migration of CoRoLa data to the new DRuKoLA environment (see below) assumes new encoding and indexing methods, mapping annotations, etc., so that the users could enjoy all the facilities of the KorAP query platform.

4.3 The Hungarian corpus

The Hungarian National Corpus is a balanced reference corpus intended to capture varieties of five selected major genres of present-day Hungarian, namely journalism, literature, (popular science), personal, and official language use. Its first version appeared in 2001 and it contained 187 million running words, morphologically annotated and tagged. The majority of the data were collected from electronic sources from within Hungary but the HNC also contains subcorpora representing Hungarian as a minority language spoken in the neighbouring countries. On the design and implementation of the first release of the corpus see Váradi (2002).

The HNC has recently been substantially upgraded and extended to gigaword size. This new release followed the original design of the corpus but the internal proportions of the genres have been changed, mainly to do justice to the ubiquitous social media. The annotation has also been overhauled and the engine and the user interface have also been modernised, employing the Manatee/Bonito framework (Rychlý, 2007). Oravecz et al. (2014) describe the corpus in more detail.

4.4 DRuKoLA

Parts of the EuReCo vision have already been implemented in the DRuKoLA-project[1], large parts of which can also be regarded as a pilot study for EuReCo (Cosma *et al.*, 2016). DRuKoLA 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). One of its main objectives is to provide a common platform for constructing various kinds of comparable corpora, based on text properties and for analysing them for contrastive linguistic purposes.

The present state of the part of DRuKoLA relevant to EuReCo is that a converter from CoRoLa-TEI-format to KorAP-XML-format has been implemented so that CoRoLa can now be accessed via KorAP. For the present moment, a large part (60%, ˜300 million words) of the textual content of CoRoLa has been incorporated as the Romanian part of the DRuKoLA content. The next step will be to fine-tune a first version of mapping functions from CoRoLa and DeReKo metadata categories to intermediate taxonomies on the basis of which virtual corpora will be dynamically generated. It seems that intermediate taxonomies for topic domains and text types will typically be necessary to arrive at sufficiently valid and fine-grained common category systems.

Romanian speech data collected in CoRoLa will be added to DRuKoLa when the appropriate processing functionality of KorAP is finalized.

4.5 DeutUng

As a second EuReCo pilot project, *DeutUng*[2] will start 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.[3]

5 Conclusions

The EuReCo initiative represents an ambitious effort of building a self-sustainable and flexible basis for comparable corpora, which is expected to offer very attractive opportunities for users but also challenges for developers. Multilinguality, which is at the root of the idea of EuReCo, together with

[1] DRuKoLA (2016-2019) 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 for 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*).

[2] DeutUng (2017-2020) is a co-operation 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.

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

the vast repositories of language data, require innovative and robust technical solutions. The cooperation of several institutions and expert groups, as envisaged by EuReCo, promises to open new research avenues in the European digital humanities. Moreover, the technical base developed in EuReCo will provide support for innovative experiments that involve linguistic resources of different types and their interconnection. Showing that a commonly agreed methodology can provide unified access to very diverse basic level linguistic representations could provide useful insights concerning linking diverse types of linguistic data (corpora, dictionaries, wordnets, etc.) and unifying access to them.

6 References

Aston, G. and Burnard, L. (1998). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Bański, P., Kupietz, M., Witt, A., Ćavar, D., Heiden, S., Aristar, A. and H. Aristar-Dry (eds.) (2012): *Proceedings of the LREC-2012 workshop on "Challenges in the management of large corpora" (CMLC-1)*. Istanbul / Paris: ELRA.

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. In: Vetulani, Z. and Uszkoreit, H. (eds.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. Poznań: Fundacja Uniwersytetu im. A., 2013: 586-587.

Bański, P., Diewald, N., Hanl, M., Kupietz, M. and A. Witt (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.

Baroni, M., Bernardini, S., Ferraresi, A., and E. Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation* 3/2009: 209-226.

CNC (2005). Czech National Corpus – SYN2005. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic.

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

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. Lit Linguist Computing (2010) 25(4): 447-464.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P. and A. Witt (2016). KorAP Architecture – Diving in the Deep Sea of Corpus Data. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portoroz / Paris: ELRA: 3586-3591.

Geyken, A. (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and Idioms*, London: 23–40.

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 the Seventh conference on International Language Resources and Evaluation (LREC'10)*: 1848-1854

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 / Paris: ELRA: 1–6.

Oravecz, Cs., Váradi, T. and Sass, B. (2014) The Hungarian Gigaword Corpus. In: Calzolari, Nicoletta et al. (eds.): *Proceedings on the Ninth International Conference in Language Resources and Evaluation (LREC'14).* Reykjavik / Paris: ELRA: 1719–1723.

Przepiórkowski, A., Górski, R. L., Łaziński, M. and P. Pęzik (2010). Recent Developments in the National Corpus of Polish. In Calzolari, N. et al. (eds.): *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris: ELRA.

Rychlý, P. 2007. Manatee/bonito–a modular corpus manager. In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno, Czech Republic: Masaryk University: 65–70.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ş. D., Boroş, T., Teodorescu, N. H., Cristea, D., Scutelnicu, A., Bolea, C., Moruz, A. and L. Pistol (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). Mannheim: IDS: 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 / Paris: ELRA.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas /Paris: ELRA: 385–389.