Originally published in: Granger, Sylviane/Lefer, Marie-Aude (Eds.): Translating and comparing languages: corpus-based insights. Selected proceedings of the fifth Using Corpora in Contrastive and Translation Studies conference. – Louvain-la-Neuve: Presses universitaires de Louvain, 2020. Pp. 257-273. (Corpora and language in use – proceedings 6)

Kupietz, Marc, Diewald, Nils, Trawiński, Beata, Cosma, Ruxandra, Cristea, Dan, Tufiş, Dan, Váradi, Tamás & Wöllstein, Angelika (2020). Recent developments in the European Reference Corpus EuReCo. In Sylviane Granger & Marie-Aude Lefer (eds) *Translating and Comparing Languages: Corpus-based Insights*. Corpora and Language in Use Proceedings 6, Louvain-la-Neuve: Presses universitaires de Louvain, 257-273.

Recent developments in the European Reference Corpus EuReCo

Marc Kupietz[1], Nils Diewald[1], Beata Trawiński[1], Ruxandra Cosma[2], Dan Cristea[4,5], Dan Tufiş[3], Tamás Váradi[6], Angelika Wöllstein[1]

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

Abstract

This paper reports on recent developments within the European Reference Corpus EuReCo, an open initiative that aims at providing and using virtual and dynamically definable comparable corpora based on existing national, reference or other large corpora. Given the well-known shortcomings of other types of multilingual corpora such as parallel/translation corpora (shining-through effects, over-normalization, simplification, etc.) or web-based comparable corpora (covering only web material), EuReCo 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.

1. Multilingual corpora for language comparison

Since the empirical turn in linguistics, corpus linguistic methods have become increasingly important not only in monolingual but also in cross-linguistic research. During the last two decades, the number of linguistic studies inspired, based on, or driven by corpus material has increased dramatically. Also, the number of corpora – both mono- and multilingual corpora – is growing rapidly. The linguist is often faced with the choice between multiple corpora of different types, and this choice has in turn consequences for outcomes of research questions and for linguistic generalizations. While the choice of a corpus as a data source for a language-specific study is naturally limited to monolingual corpora, several options regarding types of corpora are available for multilingual research. Multilingual studies can be conducted using multiple (unrelated) monolingual corpora, parallel/translation corpora or comparable corpora. Below, we will address all these possibilities and point out their advantages and shortcomings for language comparison.

1.1. Monolingual corpora

Monolingual corpora are corpora containing texts in a single language only. They are usually lemmatized and tagged for parts of speech, and are sometimes annotated for inflectional morphology, syntactic dependency and/or constituency, grammatical functions, semantic roles, named entities, anaphora and co-reference relations, information structure, etc. There are currently a large number of monolingual corpora, including specialized and reference corpora. Examples of large national reference corpora include the English language corpora American National Corpus (ANC) and British National Corpus (BNC), and the non-English language corpora DeReKo, CoRoLa and HNC, discussed in more detail below.

Monolingual corpora are characterized by very high and controlled language quality, since they contain (almost) exclusively original texts and by this, reflect language usage typical of native speakers. High linguistic quality is a very strong feature and for this reason, monolingual corpora are still frequently used in cross-linguistic research. A recent example includes a study on selectional preferences and the control behaviour of clause embedding predicates such as *try*, *promise* or *say* in German, Swedish and Dutch, conducted within the project *German Grammar in European Comparison* (GDE) at the IDS Mannheim (Hartmann *et al.* 2018). In this study, three monolingual corpora were used: DeReKo (subcorpus KoGra-DB) for German (Kupietz *et al.* 2018),

containing 4.3 G word tokens and above 170 text types, Språkbanken (subcorpus Moderna) for Swedish (Borin *et al.* 2012), containing 13.3 G word tokens and less diversified text types, and LASSY Large for Dutch (van Noord *et al.* 2006, 2013), containing 0.8 G word tokens.

As this example already indicates, a contrastive approach using monolingual corpora is faced with the problem of a low matching for size, genre, publication date and the topic of the underlying data. Due to different compositions, monolingual corpora are thus not really comparable. This poses a serious problem for contrastive research. If we represent comparability and linguistic quality as scales where the left end means low comparability or linguistic quality and the right end means high comparability or linguistic quality, then the comparability of monolingual corpora will be displayed at the very left end of the scale and linguistic quality at the right end of the scale, as shown in Figure 1.

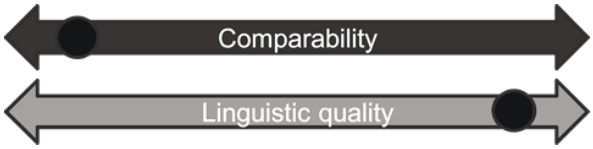

Figure 1. Low comparability and high linguistic quality in monolingual corpora

1.2. Parallel corpora

Since low comparability is a serious issue for cross-linguistic studies, many researchers prefer to use multilingual corpora, especially parallel corpora. Parallel corpora consist of original texts in one language and their translations into one or more languages. The texts are usually sentence-aligned. Due to the fact that parallel corpora provide sequences of linguistic entities (words, sentences) that convey the same meanings in the same contexts occurring in the same sorts of texts from the same periods, they can arguably serve as a perfect basis for establishing equivalence between linguistic structures (James 1980; Chesterman 1998), *i.e.* as an ideal *tertium comparationis.* Additionally, they can provide insights into similarities and differences between languages that could be overlooked when working with monolingual corpora. Because of these advantages, parallel corpora have been used as a data source in numerous contrastive studies (cf. Altenberg & Granger 2002 or Granger 2010, just to name a few examples). There is also an increasing interest in using parallel corpora in typological studies (cf. Cysouw & Wälchli 2007, among others). Finally, paral-

lel corpora play a crucial role in translational studies (cf. for example Granger *et al.* 2003), where they are often referred to as translation corpora.

Parallel corpora are used for contrastive research in many research labs worldwide, including the initiator and the coordinator of EuReCo, the IDS Mannheim. For example, in a study conducted within the project GDE at the IDS, imperatives across four languages (English, German, Polish and Czech) were investigated to validate the *agentivity hypothesis*[1] (Trawiński 2016). As a data source, the parallel corpus InterCorp (Release 6; Čermák & Rosen 2012) was used via the KonText interface. The selected parallel data included the same literary texts for each language (*e.g. 1984* by Orwell or *Le Petit Prince* by de Saint-Exupéry) and had a similar size between 1.5 and 1.7 million tokens. So, the data displayed a high degree of comparability with respect to content and size. However, by definition, they contained both original and translated texts, which poses another challenge for language comparison. As has been pointed out in the literature (*e.g.* Laviosa 1998), translated texts have specific properties that distinguish them from original texts, such as a relatively lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words over low-frequency words, a relatively greater repetition of the most frequent words and less variety in the words that are most frequently used. Baker (1995) defines the following properties typical for translated texts: *simplification* (translations tend to use simpler language), *explicitation* (translations show a tendency to spell things out) and *normalization* (translations tend to conform to the typical patterns of the target language and to overuse its features). Finally, the phenomenon of *shining-through* was identified and discussed alongside *normalization* in Teich (2003). *Shining-through* occurs when translations are oriented more towards the source language than the target language, in particular by adopting grammatical structures from the source material. To conclude, parallel corpora offer a high comparability with respect to size and content, which is a very strong feature in the context of language comparison, but the quality of the linguistic material can be lower than in monolingual corpora. This conclusion is illustrated in Figure 2.

1 The agentivity hypothesis states that imperative markers occur significantly more frequently with agentive than with non-agentive verbs.

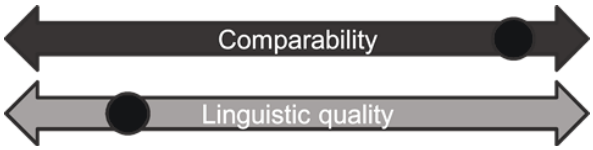

Figure 2. High comparability and lower linguistic quality in parallel corpora

1.3. Comparable corpora

As we concluded above, monolingual and parallel corpora alone are not suitable for finer-grained cross-linguistic research, because they either lack comparability or linguistic quality. A possible workaround might be to use a combination of parallel and monolingual corpora, which, however, would be complicated to handle for typical use cases. There is therefore a clear need for multilingual corpora which, on the one hand, ensure a high level of comparability in terms of content and size and, on the other hand, ensure original language quality (see Figure 3).

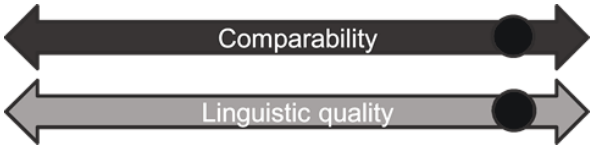

Figure 3. High comparability and high linguistic quality in an ideal multilingual corpus

Comparable corpora present an interesting option. A comparable corpus consists of two or more monolingual corpora that have similar compositions with respect to relevant properties, such as publication time, authorship, genre, topic domain, etc., and ideally contains original texts only. An early prominent example of a comparable corpus is the International Corpus of English (ICE) (Greenbaum 1991), which contains twelve corpora of different national or regional varieties of English with a controlled, similar composition. In 2017, a new international collaborative initiative on building the International Comparable Corpus (ICC) started (Kirk & Čermáková 2017). The aim of this initiative is to build many small corpora with controlled composition following the model of ICE. The primary goal is to provide highly comparable datasets for contrastive studies. The languages currently involved include Czech,

Finnish, French, German, Norwegian, Polish, Slovak, and Swedish. The ICC is an ongoing project and it is not yet available for linguistic research.

Currently, only web-based comparable corpora are available, and more specifically Aranea – Family of Comparable Gigaword Web Corpora (Benko 2014). The Aranea corpora include large corpora of controlled sizes: 1.2G words (the *Maius* edition) and 120M words (the *Minus* edition, a 10% random sample of *Maius*) and at present, they contain more than 20 languages. They were developed by means of open-source and free tools and can be used with the *NoSketch Engine* (Rychlý 2007). At the same time, the *Aranea* corpora show some drawbacks. In particular, the similarity of composition is not controlled, and moreover, it cannot be easily controlled because texts from the web notoriously lack the required metadata, such as author(s), publisher, time and place of publication, text type, topic, etc. Given this, the criterion of comparability and consequently the criterion of linguistic quality cannot be readily employed (see Figure 4).

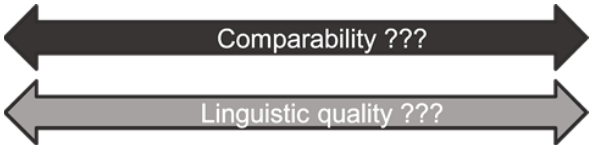

Figure 4. Uncontrolled comparability and linguistic quality in web-based comparable corpora

2. The European Reference Corpus EuReCo

The aim of the open EuReCo initiative, founded in 2013 (Kupietz *et al.* 2017), is to address the lack of high-quality multilingual comparable corpora. The idea is, however, not to build any new multilingual corpora as this would economically at least be unsustainable, but to build upon the existing monolingual reference and national corpora and to merge these virtually into tuples of comparable corpora. This means that the respective corpora remain at their locations and are networked via a common software infrastructure. The virtual merge is essential here, because the texts that national and reference corpora consist of are typically bound to their hosting institutions by license agreements that at least prohibit the copying of whole texts. This infrastructural problem is currently solved by the corpus analysis platform KorAP (Bański *et al.* 2013) which will support distributed indices, the dynamic definition of virtual subcor-

pora, and also makes the corpus data available for further linguistic analysis via a uniform interface. The construction of comparable corpora is carried out on the basis of text metadata in such a way that, ideally, the user himself can define dynamically comparable virtual subcorpora – for example by commands such as "build the largest possible corpus pair with identical composition in terms of topic, text type and year of publication". Such a dynamic definability, with the possibility of persistent storage, is important because the additional corpus and the additional requirement of comparability increase the risk of artefacts caused by corpus compositions. Unlike in monolingual corpus linguistics, not only does one corpus have to be representative of an intended language domain in relation to a research question, but also a second corpus in another language has to be comparable with that corpus. Thus, the already high risk of obtaining findings that do not say anything about the intended language domain, but are only triggered by a skewed corpus composition, is correspondingly higher when working with comparable corpora.

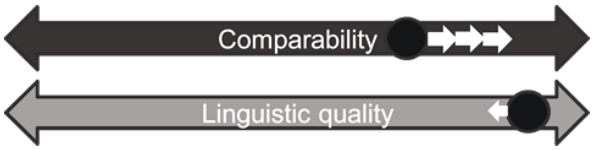

Figure 5. Gradual improvement of comparability while mostly keeping high linguistic quality by iteratively refining metadata mappings and comparability criteria

Accordingly, to be able to adjust corpus compositions, if there is any suspicion that the corpus-based findings do not reflect properties, of the language domains in question, but are rather just artefacts of skewed corpus-compositions, the construction process should ideally be iterative (see Kupietz 2015: 64) to allow for a gradual improvement of the comparability (see Figure 5) as follows:

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) seem to be artefacts of comparability criteria 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 also EuReCo's approach cannot avoid a tradeoff between comparability and corpus size.

2.1. DRuKoLA: The first EuReCo blueprint

Parts of the EuReCo vision have already been implemented in the DRuKoLA-project.[2] DRuKoLA is centered around the German Reference Corpus DeReKo (Deutsches Referenzkorpus), with more than 42 billion words (Kupietz *et al.* 2018), the largest 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, and the Reference Corpus of Contemporary Romanian Language CoRoLa (Tufiș *et al.* 2015; Barbu Mititelu *et al.* 2018), containing almost one billion words, which was publicly launched in December 2017 and can be queried via different interfaces, including KorAP (Cosma *et al.* 2016; Cristea *et al.* 2019).

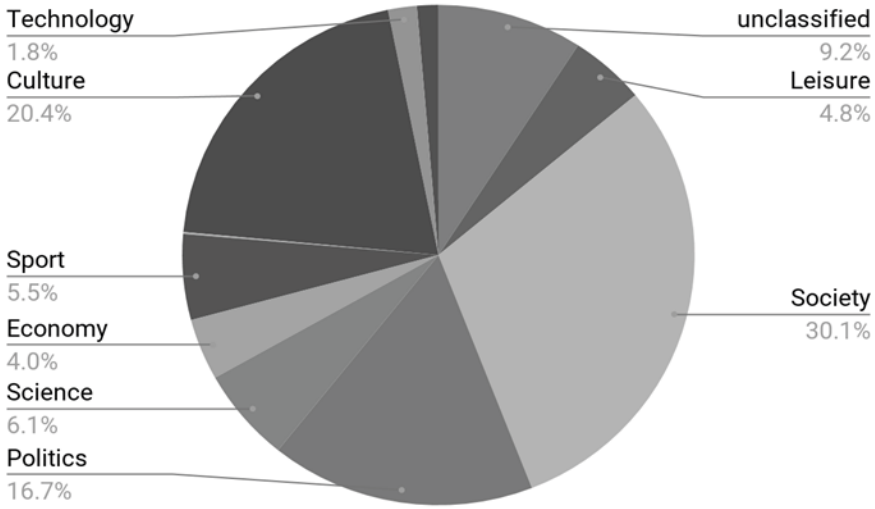

Figure 6. Number of words per DeReKo top-level topic domain in the first comparable corpus

2 DRuKoLA (2016-2018) was 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*).

The present state of the part of DRuKoLA relevant to EuReCo is that CoRoLa can be accessed publicly via KorAP and that a first virtual comparable corpus is available which is, for now, based solely on a mapping from CoRoLa's two-level topic domain taxonomy to DeReKo's topic domain taxonomy (also two-leveled, see Klosa *et al.* 2012: 88). This mapping is not yet perfect, as DeReKo uses a domain classification system based on a subset of the Open Directory (dmoz) taxonomy (see Klosa *et al.* 2012), whereas CoRoLa uses the English Wikipedia top-level domains and the Universal Decimal Classification (UDC) system (see Gîfu *et al.* 2019), resulting in slightly different categories and granularities, with, however, sufficiently similar ranges of coverage. In order to improve the mapping, the IDS plans to provide UDC and Wikipedia domains for DeReKo in the future. Thanks to the substantially larger size of DeReKo and its sufficiently similar dispersion with respect to topic domains, it was possible to build the first comparable corpus by only defining a sub-sample of DeReKo, which mimics the topic-domain composition of the whole CoRoLa, as shown in Figure 6. It has to be noted that for this first comparable German-Romanian corpus we have not controlled or thoroughly analyzed the composition with respect to publication year, and only indirectly to text type or genre via the given topic domains. A first superficial examination of the German part shows, however, that at least a large variety of genres, such as press reports, editorials, encyclopaedia articles, popular science, essays, novels, biographies, textbooks, diaries, children's books, manuals, political speeches, interviews, court decisions, letters to the editor, horoscopes, etc. (in decreasing order) are covered in the virtual DeReKo-subcorpus. Further research will show to what extent we can also make the distribution of text types and publication years comparable without the resulting comparable corpus becoming too small.

2.2. DeutUng

As a second EuReCo pilot project, *DeutUng*[3] has started to integrate the Hungarian National Corpus HNC, that has recently been substantially upgraded and extended to gigaword size (Váradi 2002; Oravecz *et al.* 2014), into EuReCo. The current state of DeutUng is that a converter for the HNC format to KorAP's input format has been developed and a first HNC sample is available via KorAP which is already being used for first pilot studies (see Section 3.2).

3 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.

3. Accessing comparable corpora with KorAP

As mentioned above, the current technical basis for EuReCo is the corpus query and analysis platform KorAP[4] (Bański *et al.* 2013; Diewald *et al.* 2016), which is currently under development at the IDS and available in public beta since May 2017. KorAP is the designated successor of the corpus search and management system COSMAS II[5] as the main access point to DeReKo. KorAP's design aims to be independent regarding research questions and underlying data, and is therefore adaptable for corpora in different languages with different annotations. It also supports multiple corpus query languages (*e.g.* Poliqarp, COSMAS II QL, AnnisQL), welcoming users with varying expertise regarding corpus analysis tools. For comparable corpora in the EuReCo scenario, KorAP provides some essential features, in particular

- its ability to manage corpora that are physically located at different places, in order to comply with typical license restrictions (Kupietz *et al.* 2014);
- its ability to dynamically create virtual subcorpora based on text properties and to manage these virtual corpora in a persistent way, for example allow for reusability and reproducibility.

3.1. Accessing the German-Romanian Comparable Corpus

A subcorpus of DeReKo, comparable in size and composition, was compiled for CoRoLa, based on metadata information and document metrics. This subcorpus is stored as a persistent virtual corpus (VC) in KorAP and can be referenced[6] (optionally as part of a more complex VC) to restrict search and analysis to all documents in the comparable corpus. The German-Romanian comparable corpus currently consists of more than 3 million documents, comprising 940 million word tokens (see Figure 7). Although metadata and annotations differ, both corpora can be searched in a comparable way in KorAP. Figure 8 shows, for example, a query for postnominal adjective sequences conducted in both corpora, with the match count indicating a more common postnominal pattern in Romanian, motivating further research in the structure of these patterns. Romanian is a language that allows both pre- and postnominal positions for

4 https://korap.ids-mannheim.de/

5 https://cosmas2.ids-mannheim.de/

6 The reference identifier is "drukola.20180909.1b_words".

adjectives (sometimes even simultaneously) while in German the prenominal position is the regular case.

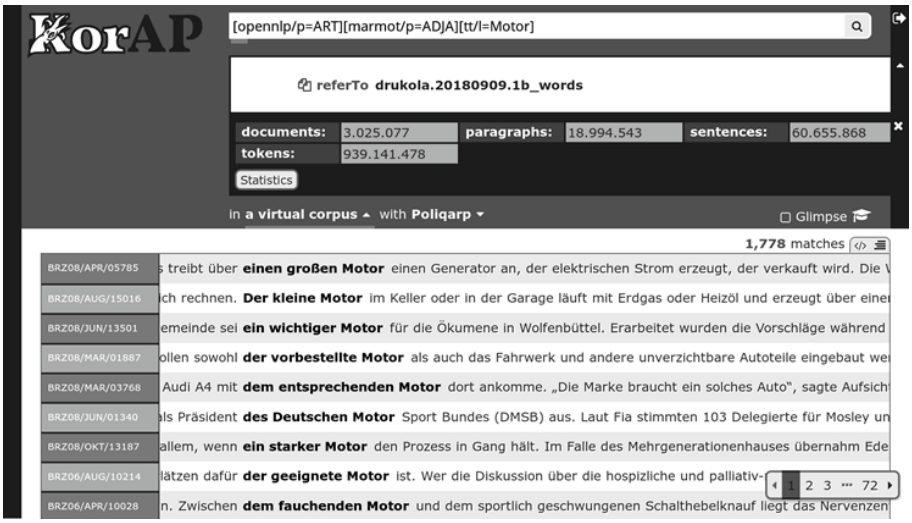

Figure 7. Referring to a persistent virtual corpus in KorAP

Figure 8. Searching the comparable corpus of German and Romanian in KorAP for a sequence of a noun followed by two adjectives, expressed in Poliqarp QL and referring to different underlying annotations

An in-depth study may then compare these different patterns regarding adjective positions in both corpora by refining the queries to recognize language specific annotations (cf. Cornilescu & Cosma 2019).

3.2. Accessing the German-Hungarian comparable corpus

Within the context of the DeutUng project, first portions of the HNC have been integrated into EuReCo and small German-Hungarian comparable corpora are already available for querying with KorAP.

One of the research questions addressed in the DeutUng project is the distribution of pronouns as correlatives to complement clauses (Hartmann *et al.* 2017). In Hungarian, the correlative pronoun *azt* is possible in structures headed by assertive verbs (such as *say*) but it is not possible in structures headed by factive verbs (such as *regret*). In German, exactly the opposite is the case: The pronoun *es* can be used in complex sentences with factive verbs but it cannot with assertive verbs. This is illustrated by the examples (1-4), taken from Molnár (2015: 211-212).

(1) Péter *azt* **mondta**, hogy gyakran találkoznak munka után. **assertive (HU)**
Peter it-ACC said-3SG that often gather-3PL work after
'Peter said that they often meet up after work'

(2) Péter *(*azt)* **bánja**, hogy elfogadta a meghívást. **factive (HU)**
Peter it-ACC regrets that accepted-3SG the invitation-ACC
'Peter regrets that he has accepted the invitation'

(3) Peter **behauptet** *(*es)*, dass sie sich […] oft treffen. **assertive (DE)**
Peter claims it-ACC that they REFL […] often gather-3PL
'Peter claims that they often meet up […].'

(4) Peter **bedauert** *es*, dass er die Einladung angenommen hat. **factive (DE)**
Peter regrets it-ACC that he the invitation-ACC accepted has
'Peter regrets that he has accepted the invitation'

However, as already pointed out in Molnár (2015), among others, in some contexts or under specific circumstances (such as focus), the Hungarian correlative *azt* seems to be possible with factive predicates as well. The usage of the Ger-

man *es* in different (information structural) conditions does not yield a clear picture, either. The goal of the contrastive analysis envisaged in the DeutUng project is to identify the factors determining the distribution of these pronouns in their correlative function in these two languages. Figure 9 displays partial results of searching DeReKo and HNC via KorAP for correlative pronouns in German and Hungarian with factive and assertive verbs, respectively.

Figure 9. Searching DeReKo and HNC for correlative pronouns with factive and assertive verbs. The different highlighting of verb types indicates a reversed usage of the pattern.

4. Conclusions and outlook

We have shown how the EuReCo initiative addresses the current lack of multilingual corpora that satisfy both the criterion of high linguistic quality, including size and diversity, and the criterion of comparability. We also show how this can be done in an economically feasible way, by building upon and re-using existing corpora and joining them virtually, using the corpus query platform KorAP. In addition, we have sketched EuReCo's approach to tackle the complex and error-prone definition of comparability by iteratively adjusting the comparability criteria. Finally, we have demonstrated how the general approaches are already being applied using KorAP in contrastive studies that compare German with Romanian and Hungarian in the two EuReCo pilot projects DRuKoLA and DeutUng.

The next steps will be to improve DeReKo's topic domain classification, to further integrate the HNC into EuReCo, to work on special KorAP features for comparable corpora, and to iteratively test and improve the first German-Romanian corpus, based on quantitative and qualitative case studies. Furthermore, EuReCo is happy to welcome corpora for additional languages.

References

Altenberg, B. & Granger, S. (eds) (2002). *Lexis in Contrast. Corpus-based Approaches*. Amsterdam & Philadelphia: Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research, *Target* 7(2), 223-243.

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 (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. Mickiewicza, 586-587.

Barbu Mititelu, V., Tufiș, D. & Irimia, E. (2018). The reference corpus of the Contemporary Romanian Language (CoRoLa). In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, S. Sara Goggi & H. Mazo (eds) *Proceedings of LREC 2018*. Miyazaki & Paris: ELRA, 1178-1185.

Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In P. Sjka, A. Horák, I. Kopeček & K. Pala (eds) *Text, Speech and Dialogue. TSD 2014. Lecture Notes in Computer Science,* vol. 8655. Cham (Switzerland): Springer.

Borin, L., Forsberg, M. & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In N. Calzolari, K. Choukri, T. Declerck, M. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds) *Proceedings of LREC 2012*. Istanbul & Paris: ELRA, 474-478.

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 17(3), 411-427.

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

Cornilescu, A. & Cosma, R. (2019). Linearization of attributive adjectives in Romanian and German. *Revue roumaine de linguistique* 3, 307-322.

Cosma, R., Cristea, D., Kupietz, M., Tufiş, D. & Witt, A. (2016). DRuKoLA – Towards contrastive German-Romanian research based on comparable corpora. In P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen & A. Witt (eds) 4th Workshop on Challenges in the Management of Large Corpora. *Proceedings of LREC 2016.* Portorož & Paris: ELRA, 28-32.

Cristea, D., Diewald, N., Haja, G., Mărănduc, C., Barbu Mititelu, V., Onofrei, M. (2019). How to find a shining needle in the haystack. Querying CoRoLa: solutions and perspectives, *Revue roumaine de linguistique* 3, 279-292.

Cysouw, M. & Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *STUF - 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 LREC 2016*, Portorož & Paris: ELRA, 3586-3591.

Gîfu, D., Moruz, A., Bolea, C., Bibiri, A. & Mitrofan, M. (2019). The methodology of building CoRoLa. *Revue roumaine de linguistique* 3, 241-253.

Granger, S. (2010). Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Contemporary Foreign Language Studies* 2 (Shanghai Jiao Tong University), 14-21.

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

Greenbaum, S. (1991). The development of the international corpus of English. In K. Aijmer & B. Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 83-92.

Hartmann, J. M., Mucha, A., Trawiński, B. & Wöllstein, A. (2018). Selectional preferences for (non-)finite structures as indicators of control relations: A cross-Germanic corpus study. Paper presented at the *Grammar and Corpora Conference 2018*, 15-17 November, University of Paris-Diderot (France).

Hartmann, J. M., Schlotthauer, S., Trawiński, B. & Wöllstein, A. (2017). Sprachvergleich: Einblicke in die aktuelle kontrastive Forschung am IDS: Nominal- und Verbgrammatik. Paper presented at the Kick-off of the DeutUng project, 19 October 2017, University of Szeged (Hungary).

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

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: IDS, 7-12.

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

Kupietz, M. (2015). Constructing a corpus. In Ph. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, 62-75.

Kupietz, M., Lüngen, H., Bański, P. & 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)*. Reykjavik & Paris: ELRA, 1-6.

Kupietz, M., Lüngen, H., Kamocki, P. & Witt, A. (2018). The German reference corpus DeReKo: New developments – new opportunities. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, S. Sara Goggi & H. Mazo (eds) *Proceedings of LREC 2018*. Miyazaki / Paris: ELRA, 4353-4360.

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: IDS, 15-19.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4), 557-570.

Molnár, V. (2015). The predicationality hypothesis. The case of Hungarian and German. In K. É. Kiss, B. Surányi & É. Dékány (eds) *Approaches to Hungarian* 14. Papers from the 2013 Piliscsaba Conference. Amsterdam & Philadelphia: Benjamins, 209-244.

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 of LREC 2014.* Reykjavik & Paris: ELRA, 1719-1723.

Rychlý, P. (2007). Manatee / Bonito - A modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 65-70.

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

Trawiński, B. (2016). Messung der Distanz zwischen grammatischen Kategorien im sprachübergreifenden Kontext. In A. V. Averina (ed.). *Grammatitscheskije kategorii v kontrastivnom aspektje. Sbornik nautschnych statjej no materialam mjeschdunarodnoj konfjerjentschii*, Moskva, Volume 1. Moscow: Moscow Pedagogical University, 116-120.

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: IDS, 5-10.

van Noord, G., Bouma, G, Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Tjong Kim Sang, E. & Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: Lassy. In P. Spyns & J. Odijk (eds) *Essential Speech and Language Technology for Dutch.* Heidelberg, New York, Dordrecht & London: Springer, 147-164.

van Noord, G., Schuurman, I. & Vandeghinste, V. (2006). Syntactic annotation of large corpora in STEVIN. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (eds) *Proceedings of LREC 2006*. Genoa & Paris: ELRA, 1811-1814.

Váradi, T. (2002). The Hungarian national corpus. In M. Rodríguez & C. Araujo (eds) *Proceedings of LREC 2002*, Las Palmas & Paris: ELRA, 385-389.