EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research

Marc Kupietz, Piotr Bański, Nils Diewald, Beata Trawiński, Andreas Witt
Leibniz Institute for the German Language (IDS)
R5 6-13, 68161 Mannheim, Germany
{kupietz, banski, diewald, trawinski, witt}@ids-mannheim.de

Abstract

This paper gives an overview of recent developments concerning the European Reference Corpus EuReCo, an open long-term initiative aimed at providing and using virtual and dynamically definable comparable corpora based on existing national, reference or other large corpora. Given the problems and shortcomings of other types of multilingual corpora – such as the shining-through effects in parallel corpora or the limitation to web material only in web-based comparable corpora – EuReCo constitutes a unique linguistic resource that offers new perspectives for fine-grained cross-linguistic research. The approach advocated here puts forward new solutions to notorious IPR and licensing issues, as well as to challenges of interoperability. It also addresses methodological questions concerning comparability and representativeness. While the focus of this paper is on EuReCo's implementation-based approach to ensuring interoperability in a feasible and maintainable way, it also presents preliminary results of pilot comparative studies on light verb constructions in German, Romanian, Hungarian, Polish and Bulgarian, and reports on recent extensions and plans.

Keywords: Reference Corpora, National Corpora, Federated Corpora, Multilingual Corpora, Cross-Linguistic Research, Comparability

1. Introduction

The challenge of comparability in multilingual studies relates both to the language data itself and to the methods applied. In this paper, we discuss the relevant features of the available corpus types from a linguistic perspective and point out their advantages and disadvantages, particularly for cross-linguistic research (Section 2). Against this background, we present an approach to using comparable corpora without having to build them: the European Reference Corpus EuReCo. EuReCo is an open long-term initiative that aims at providing and using virtual and dynamically definable comparable corpora based on existing national, reference or other large corpora. Section 3 presents the basic ideas behind EuReCo and the previous work. Section 4 introduces and discusses access to federated corpora and EuReCo's approach to interoperability, with the corpus analysis platform KorAP as a working implementation, and Section 5 presents recent developments within the EuReCo initiative, including applications in the area of cross-lingual studies of light verb constructions (Section 5.4). Section 6 summarizes the paper and sketches the next steps.

2. State of the Art

From the linguistic point of view, there exist several advantages and disadvantages of monolingual corpora, parallel corpora and the available comparable corpora. Based on Kupietz et al. (2020b) and Trawiński and Kupietz (2021), we argue that there is a great need in cross-linguistic research for high-quality multilingual data whose degree and angle of comparability can be flexibly adjusted.

2.1. Monolingual Corpora

Monolingual corpora are, by definition, corpora that contain texts in a single language. They are characterized by a very high and controlled linguistic quality, as they typically contain (ideally only) original texts and thus reflect native language usage. There is currently a large number of monolingual corpora, including both (mostly smaller) specialized corpora and national or reference and other very large general corpora, such as the British National Corpus (BNC; Aston and Burnard, 1998; Brezina et al., 2018), the Corpus of Contemporary American English (COCA; Davies, 2011), the Czech National Corpus (CNC; Křen, 2020), the Romanian Contemporary Language Reference Corpus (CoRoLa; Barbu Mititelu et al., 2018), the German Reference Corpus (DeReKo; Kupietz et al., 2010, 2018), the Hungarian National Corpus (HNC; Váradi, 2002; Oravecz et al., 2014), and the Polish National Corpus (NKJP; Przepiórkowski et al., 2012) — of which the last four are already, at least partially, integrated into EuReCo.[1]

[1] Numerous corpora, both monolingual and multilingual, are also provided by Sketch Engine (see, e.g., Kovář et al., 2016, https://www.sketchengine.eu), but they are not freely available to the full extent.

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 94–103

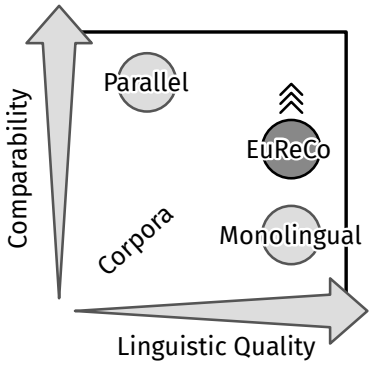

Figure 1: Comparability and Linguistic Quality

The high linguistic quality of monolingual corpora is one of the main reasons why they are used not only for single-language studies but also for cross-language research, both as a source of evidence and for advanced quantitative analyses (see, for example, the numerous contributions in Trawiński et al., 2023). However, while the high linguistic quality of monolingual corpora is a major advantage, their use as a basis of data for cross-linguistic research has obvious shortcomings, leading to the question of whether and to what extent the results of studies performed on different languages are comparable with one another. This is due to the large differences between the individual monolingual corpora in terms of size, composition and annotation (see, e.g., Kupietz et al., 2020b; Trawiński and Kupietz, 2021).

Since the low comparability of monolingual corpora (despite their high linguistic quality as illustrated in Fig. 1) poses a serious empirical and methodological problem for language comparison, multilingual corpora, and especially parallel corpora, which are discussed in the following section, are predominantly used in cross-linguistic research.

2.2. Parallel Corpora

Parallel corpora consist of original texts in one language (source language) and their translations in other languages (target languages), which is why they are sometimes called translation corpora (e.g. in translation studies). The parallel texts are usually aligned at sentence level and are sometimes linguistically annotated. There are now a number of electronic parallel corpora that are freely accessible and can be searched using various web-based research and analysis systems. Among the largest and most popular are currently The Open Parallel Corpus OPUS (Tiedemann and Nygaard, 2004; Tiedemann, 2012), the multilingual parallel corpus InterCorp (Čermák and Rosen, 2012; Rosen et al., 2019), and The European Parliament Proceedings Parallel Corpus Europarl (Koehn, 2005). In addition, there exist numerous smaller parallel corpora, which are either bilingual or consist of only a few languages, but often contain more detailed and accurate linguistic information due to (partly) manual annotation. Examples are the Stockholm MULtilingual TReebank SMULTRON (Volk et al., 2015) or the the CroCo corpus (Hansen-Schirra et al., 2006).

Parallel data, as provided by parallel corpora, represent linguistic units (words, phrases, sentences) in two or more languages that are translation equivalents of each other (based on functional equivalence) and as such convey the same (or similar) meanings. It is also important that these linguistic units can be viewed in-context in the respective source and target languages and within the same text types in relation to exactly the same topics, time periods, etc. Because of these properties, parallel data provide a perfect basis for determining functional equivalence between linguistic structures in a cross-linguistic context. In other words, they can be used as a perfect *tertium comparationis* (see also James, 1980; Chesterman, 1998). In addition, parallel data provide insights into cross-linguistic similarities and divergences that can easily be overlooked when working with monolingual corpora. These properties of parallel data have been recognized early in cross-linguistic research and have been utilized in numerous studies in the fields of contrastive linguistics (see, e.g., Altenberg and Granger, 2002; Granger, 2010; Trawiński et al., 2023), language typology (see Cysouw and Wälchli, 2007, and other articles in the containing volume) and translation studies (see, e.g., Granger et al., 2003; Granger and Lefer, 2022).

However, despite the high degree of comparability in terms of content and size, parallel corpora provide a relatively small and undifferentiated database. In general, the more languages are used for comparison, the more the number and differentiation of parallel texts decreases. In addition, there is often a strong disproportion of original texts as opposed to translated texts (cf. the discussion in Kupietz et al., 2020b; Trawiński and Kupietz, 2021).

Due to their special properties, translation texts are considered as a *third code*, i.e. a special type of text that differs from both the source language and the target language (cf. Frawley, 1992; Baker, 1993). Baker (1995) observes that translations tend to use simpler language (*simplification*), to clarify things (*explication*), and to overuse typical patterns of the target language (*normalization*). Laviosa (1998) further identifies the following properties of translated texts: relatively low proportion of lexical words compared to functional words, relatively high proportion of high-frequency words compared to low-frequency words, frequent repetition of frequent words, and low variety of frequent words. In addition to *normalization*, Teich (2003) defines and

investigates the phenomenon of *shining-through* empirically on the basis of German-English and English-German corpora, using various grammatical constructions (such as passive and relative clauses) as examples. *Shining-through* occurs when translations are closer to the source language than to the target language. *Normalization* in terms of Teich (2003) occurs when translations are more closely oriented to the target language than would be expected.

To conclude, parallel corpora are highly comparable in terms of size and content, which is crucial for language comparison. In contrast, the quality of the linguistic material is poor compared to monolingual corpora (see Fig. 1).

2.3. Comparable Corpora

As explained above, monolingual and parallel corpora alone are suitable for contrastive linguistic research of finer granularity only to a limited extent, since, in short, they lack either comparability or linguistic quality. One way to avoid these limitations is to combine the parallel or monolingual corpora in question and to form hypotheses based on the parallel corpora, and afterwards to test them against the monolingual corpora. The disadvantage of this approach, however, is that it is time-consuming. This disadvantage can be decisive, especially in a corpus-led, explorative approach, where it is important to derive the most promising hypotheses and test them quickly in order to ultimately gain linguistic knowledge. In order to be able to assess the comparability and generalizability of corpus findings, further manual and argumentative work is also necessary. The situation is even more difficult if the corpora are only used indirectly via a language model in distributional analyses. It would therefore be better in most cases to be able to start from comparable corpora (McEnery and Xiao, 2007) of high quality.

To our knowledge, the only available comparable corpus with a broader coverage spectrum is Aranea – the family of comparable Gigaword web corpora (Benko, 2016). Aranea contains corpora of more than 20 languages, including corpora of German from Switzerland and from Austria, with controlled sizes of 120M and 1.2G words respectively. They can be queried online using the NoSketch engine (Rychlý, 2007) or KonText (Machálek, 2020). However, their limitation is that the comparability of the composition is not controlled and cannot be easily verified, since the Aranea corpora are fed exclusively from web texts that do not systematically contain the necessary metadata.

3. The European Reference Corpus EuReCo

3.1. Basic Assumptions

The European Reference Corpus EuReCo (Kupietz et al., 2017) is an open initiative founded around 2012 by the Leibniz Institute for the German Language (IDS) and the Academies of Sciences in Poland, Romania and Hungary. EuReCo is based on two fundamental assumptions. First, the creation of a significant number of new comparable corpora in Europe is unlikely to be feasible in the foreseeable future, also for reasons concerning research funding policy. The idea of EuReCo was therefore from the outset not to create new corpora, but rather to draw exclusively on the existing national and reference corpora, this way ensuring sufficient size and high linguistic quality. The second fundamental assumption of EuReCo is that general comparability of corpora is not an achievable and therefore not a particularly sensible goal (Kupietz and Trawiński, 2022).

EuReCo follows an approach that is complementary to the International Comparable Corpus (ICC) initiative (Kirk and Čermáková, 2017; Čermáková et al., 2021; Kupietz et al., 2023), which uses small corpora of predefined composition. In contrast to the ICC, no static extracts are copied from the source corpora of EuReCo – instead, the entire relevant corpora are linked virtually by means of the appropriate research software. Four reasons motivated this decision: firstly, this seemed to be the only way to fundamentally solve future copyright and licensing problems; secondly, it ensured that EuReCo would automatically benefit from future extensions of the corpora involved; thirdly, it seemed essential to use a common research platform anyway and to distribute its further development and maintenance across as many shoulders as possible. The fourth reason is the failure to establish a universal set of criteria for general comparability of corpora.

3.2. Comparability and Representativeness

Kupietz and Trawiński (2022) point out that corpora of reasonable size and diversity cannot in general be perfectly comparable, as there will always be some criterion by which the corpora differ. Whether an uneven distribution of a variable is relevant depends on the specific question being asked. Moreover, also monolingual corpora cannot be generally representative either, since their population (=language) cannot be generally defined (Evert, 2006; Koplenig, 2017). Thus, whether a pair of corpora is *sufficiently* comparable and representative cannot be decided a priori, but depends

on the research question and the target language domain. For these reasons, a *primordial sample* approach (Kupietz et al., 2010) was chosen for EuReCo. This approach, which has been used since the 1990s for the German Reference Corpus (Teubert, 1998), invites users to use either predefined (comparable) virtual corpora or to define suitably representative and comparable corpora for the respective research question on the basis of metadata, roughly in accordance with stratified sampling. This construction of virtual comparable corpora can typically be understood as an iterative optimization process (Cosma et al., 2016). First, subcorpora are sampled from the monolingual corpora in such a way that they have similar text / token distributions with respect to relevant metadata variables, such as subject area, text type, year of publication. Then the investigations are carried out and the virtual comparable corpus definitions (or, if necessary, the research hypotheses) are iteratively refined until it can be ruled out that the findings are only due to inadequate comparability criteria or other confounding factors or artifacts. In this way, the comparability of the corpora can be effectively optimized specifically for individual research questions, as sketched in Fig. 1 (see Kupietz, 2015, for a more comprehensive description).

3.3. Previous Work

The idea of reusing existing large corpora and making subsets of them comparable is not new and, as far as we know, was first attempted by Bekavac, Osenova, Simov, and Tadić (2004), who built a Bulgarian-Croatian comparable corpus on the basis of two newspaper subcorpora from larger reference corpora of Bulgarian and Croatian.

As part of the EuReCo initiative, two large pilot projects have been carried out so far: DRuKoLA (2016–2018) and DeutUng (2017–2021)[2]. As part of DRuKoLA, the Contemporary Reference Corpus of the Romanian Language CoRoLa (Barbu Mititelu et al., 2018; Tufiș et al., 2019) was made searchable via KorAP (Bánski et al., 2013).[3] In addition, the first German-Romanian comparable corpora were defined in the project. For these, only the topic domain variable was controlled, and a random sample was drawn from DeReKo so that it contains the same token and text quantities as CoRoLa for each topic domain (see Kupietz et al., 2020b, for details). A corresponding virtual subcorpus of DeReKo also has a very similar token distribution with regard to the year of publication (Trawiński and Kupietz, 2021, p. 223) and can be publicly queried via KorAP.[4] Several smaller pilot studies have also already been conducted on the basis of the German-Romanian comparable corpora (Kupietz et al., 2020b).

As part of the DeutUng project, the Hungarian National Corpus HNC with over one billion words was made searchable via KorAP.[5] Individual small contrastive studies were also carried out.

4. Access to Federated Corpora for Cross-Linguistic Research

The use of already existing, large national or reference corpora for cross-linguistic studies means that, on the one hand, the rights to the data are held by separate institutions and therefore data cannot be provided centrally by a single instance (especially for legal reasons; see Fig. 2a). On the other hand, the use of different corpus analysis platforms provided by these institutions (with different feature sets, different frontends, and different API methods for accessing the separate corpus data) means reduced methodological comparability and increased demands placed on the user's skills when it comes to operating multiple systems (see Fig. 2b). A technical solution to access these corpora for contrastive research must therefore offer both geographical distribution of the data, and parallel searchability and analyzability using comparable methods.

4.1. Specification-Based vs. Implementation-Based Interoperability

In recent years, the CLARIN Federated Content Search[6] (FCS; Trippel, 2013) has proven to be the most important technical initiative for decentralized cross-linguistic research. The FCS specifies protocols and formats that corpus providers have to implement in order to make their data accessible for comparison (see Fig. 2c). This form of specification-based interoperability (comparable to other Internet specifications such as HTML or email) has some advantages in a heterogeneous corpus landscape. The most prominent advantage is certainly the autonomy of the data providers, who can decide to what degree they want to be interoperable and who can provide not only existing corpora but also ex-

[2]Both funded by the Alexander von Humboldt Foundation as Institute Partnerships

[3]See https://korap.racai.ro/

[4]The following link leads to a modifiable search within a predefined virtual DeReKo subcorpus, which is comparable to CoRoLa in terms of topic domain composition: https://korap.ids-mannheim.de/?q=%3Cbase/s=t%3E&cq=referTo%20drukola.20180909.1b_words

[5]See https://korap.nlp.nytud.hu/

[6]https://contentsearch.clarin.eu/

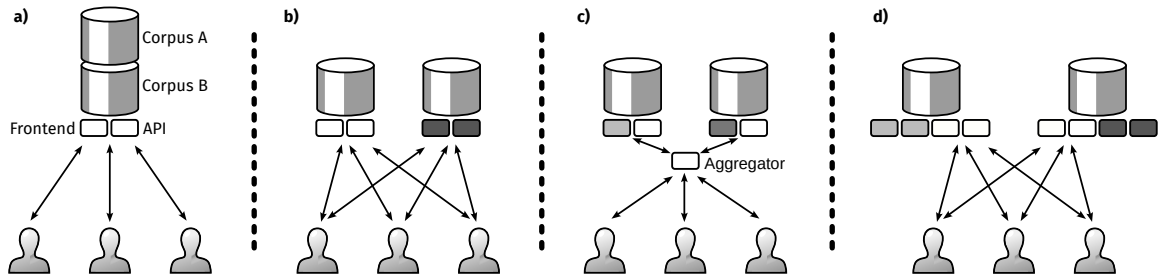

Figure 2: Querying comparable corpora: a) provided by a central instance; b) provided by different instances and interfaces; c) provided by different instances but comparable interfaces; d) provided by different instances but identical interfaces

isting corpus analysis platforms, optimized for their data and their users.

However, specification-based interoperability also has some disadvantages that can be a hindrance to the primary application scenario of EuReCo, namely to allow detailed language comparison studies:

- The scope of features provided is limited to the intersection of the feature sets provided by all participating systems (and is therefore often pretty basic);
- Innovations in the specification to extend or adapt the scope of features require new implementations and maintenance work at multiple locations and can only be used once this work has been carried out on all participating systems.

For this reason, an implementation-based approach to interoperability was chosen for EuReCo (comparable to, e.g., Shibboleth[7]), in which corpus providers deploy a special platform that is developed openly and can be used in parallel with existing corpus analysis systems (see Fig. 2d) with as little maintenance and cost as possible.

4.2. KorAP as a Tool for Implementation-Based Interoperability

While it has yet to be decided which software solution (or solutions) are going to be used for EuReCo in the future, the corpus search and analysis platform KorAP has been applied for the previous pilot projects. KorAP (Bański et al., 2012; Diewald et al., 2016) has initially been developed primarily as an access point to DeReKo, but is suitable for most corpora[8] with arbitrary metadata and arbitrary annotations due to its agnostic approach regarding data and research questions. KorAP is also under active development as part of a standing project at IDS Mannheim, and is adaptable to various usage scenarios (e.g. with localization, plugins, an extensible set of query languages, and due to its open development[9]). Corpora that conform to the TEI-P5 guidelines (TEI Consortium, 2024) can be converted to the required target format and enriched with annotations in the CoNLL-U format[10] (which is supported by numerous existing annotation tools) using an open source conversion pipeline.[11]

The supported definition of virtual corpora based on metadata (Kupietz et al., 2010) makes it possible to create sub-corpora for search and analysis that can be referenced beyond instances according to certain criteria and are thus comparable in a decentralized scenario (as *virtual collections*; cf. Broeder et al., 2008).[12] KorAP also supports a complex and finely granular rights management system, which gives corpus providers exclusive control over the data to be made available, even in the case of decentralized access.

5. Recent Developments

5.1. Addition of Further Languages

This section reports on the steps taken towards two planned extensions to the coverage of EuReCo.

5.1.1. National Corpus of Polish

A pilot conversion project from a 1M sample of the National Corpus of Polish (NKJP; Przepiórkowski

[7] https://www.shibboleth.net/

[8] Restrictions may concern, e.g., word segmentations.

[9] https://github.com/KorAP/; provided under a BSD-2-clause License

[10] https://universaldependencies.org/format.html

[11] See, e.g., https://github.com/KorAP/KorAP-XML-TEI and https://github.com/KorAP/KorAP-XML-CoNLL-U

[12] This sampling procedure, as described in Sec. 3.2, can already be implemented using KorAP. However, the API interface or the R (Kupietz et al., 2020a) or Python libraries (Kupietz et al., 2022) are still required for downsampling parts of the defined virtual corpora to their intended sizes.

et al., 2012) to the native format of KorAP was successfully concluded in the autumn of 2023. The project targeted a dataset published in May of that year as part of the Morfeusz test data suite[13], referred to as NKJP-SGJP, where the latter part of the name stands for "grammatical dictionary of the Polish language". This dataset is based on the original NKJP1M v. 1.2, published under CC-BY, and includes a format modification in the morphological layer that makes it more suitable for mass conversion. The additional advantage is that this version receives, on a nearly monthly basis, manual improvements of the POS and morphosyntactic annotation, according to the tagset defined for the Morfeusz tagger (Woliński, 2014), which is currently a *de-facto* standard tagger for numerous projects developing Polish language resources.

The converted tagset, apart from a layer of morphosyntactic annotation and NER information, includes also information on all possible morphological parses of its segments, before the phase of morphosyntactic disambiguation. This makes it possible to test a potential extension to the Poliqarp+ parser used in KorAP in order to handle the ~-operators (Janus and Przepiórkowski, 2007).

5.1.2. Bulgarian National Reference Corpus

Spassova (2023) has adapted the Bulgarian National Reference Corpus (BNRC; Simov et al., 2004) for use with KorAP and carried out a pilot comparative study. However, the metadata for the BNRC has not yet been mapped, and it is not yet publicly available for querying.

5.2. EuReCo as a CLARIN Project

At the EuReCo Kick-Off Workshop held on 18 October as part of the CLARIN Annual Conference 2023, the ideas underlying EuReCo were discussed with 26 invited representatives of different countries, regions and languages.

The main topics of discussion were the clarification and viability of the EuReCo solution for IPR and licensing issues, the challenge of metadata mapping, and the implementation-based approach to solving the interoperability problem with its additional costs of data conversion and of setting up and maintaining an additional corpus analysis tool. Following the discussion, which also touched on the issue of desirability versus feasibility, the final, unanimous decision was to propose a new joint CLARIN project to implement EuReCo.

[13] http://morfeusz.sgjp.pl/download/

5.3. Harmonization of Text Classification Metadata

The biggest challenge for the EuReCo approach is that the existing text type and domain classification systems differ among the national and reference corpora, so that these must either be mapped to a common taxonomy or to each other.

To address the issue of common domain classification, we are currently experimenting with fine-tuning multilingual Large Language Models using the English Wikipedia top-level domain as well as the standard library domain classification systems established in the Dewey Decimal Classification (DDC) and the Universal Decimal Classification (UDC).

5.4. Ongoing Work on Light Verb Constructions

Ongoing contrastive linguistic applications of EuReCo focus on comparisons of syntagmatic patterns in German with Romanian, Hungarian and Polish, and their variation depending on the context. Inspired by an approach by Taborek (2020), collocation analyses have been carried out in order to explore light verb constructions and their variation depending on text-external variables (text type, topic domain). These studies also serve to evaluate the properties of the respective comparable corpus definitions, KorAP's support for contrastive analyses[14], and the viability of the EuReCo approach, in general.

So far, the individual results of these studies were not particularly surprising. However, the overall results were surprising in that they supported our assumptions to a greater extent than we had anticipated.

The studies show, for example, that the results of collocation analyses vary greatly with the composition of the corpus and are particularly dependent on the proportion of certain topic domains (see Kupietz and Trawiński, 2022, p. 429ff). The type and strength of the effects differ depending on the language and on the light verb constructions analyzed. The richness of the results and the strong dependence on the composition of the comparable corpora show that even simple lexicological-syntagmatic analyses benefit greatly from an approach that allows for the dynamic definition of (comparable) corpora. Furthermore, the pilot studies, including those using the ICC, have also shown that corpora (samples) with a size of 1 million words or less are not sufficient for the study of even relatively frequent light verb constructions (Bański

[14] Contrastive collocation analyses are not yet possible via the KorAP web interface. Instead, we used KorAP's R library. This also facilitated replication when analyzing the effects of different corpus compositions.

et al., 2023; Kupietz et al., 2023), so that the size of national and reference corpora, with several 100 million words, seems to be a good minimum for conducting finer-grained cross-linguistic research.

6. Conclusions and Outlook

The provision of comparable corpora for cross-linguistic research is associated with scientific, technical, legal and sometimes political challenges. With an implementation-based model for federated access to these corpora, we are pursuing an approach that is as cost-effective and low-maintenance as possible while still ensuring a high level of variability and methodological rigor.

In the next steps, further national and reference corpora are going to be integrated into EuReCo. Meanwhile, different approaches to mapping metadata (in particular topic domain and text type) to common classification systems are going to be evaluated.

7. References

Bengt Altenberg and Sylviane Granger, editors. 2002. *Lexis in Contrast: Corpus-based approaches*, volume 7 of *Studies in Corpus Linguistics*. Benjamins, Amsterdam.

Guy Aston and Lou Burnard. 1998. *The BNC Handbook*. Edinburgh University Press.

Mona Baker. 1993. Corpus linguistics and translation studies – Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In honour of John Sinclair*, pages 233–250. Benjamins, Amsterdam.

Mona Baker. 1995. Corpora in Translation Studies: An overview and some suggestions for future research. *Target*, 7(2):223–243.

Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1178–1185, Miyazaki, Japan. European Language Resources Association (ELRA).

Piotr Bański, Nils Diewald, Marc Kupietz, and Beata Trawiński. 2023. Applying the newly extended European reference corpus EuReCo. Pilot studies of light-verb constructions in German, Romanian, Hungarian and Polish. In *Book of Abstracts of the 10th International Contrastive Linguistics Conference (ICLC-10), 18-21 July, 2023, Mannheim, Germany*, pages 274–276, Mannheim. IDS-Verlag.

Piotr Bański, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld, and Andreas Witt. 2012. The New IDS Corpus Analysis Platform: Challenges and Prospects. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2905–2911, Istanbul, Turkey. European Language Resources Association (ELRA).

Božo Bekavac, Petya Osenova, Kiril Simov, and Marko Tadić. 2004. Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Vladimír Benko. 2016. Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4245–4248, Portorož, Slovenia. European Language Resources Association (ELRA).

Vaclav Brezina, Robbie Love, and Karin Aijmer, editors. 2018. *Corpus Approaches to Contemporary British Speech: Sociolinguistic studies of the Spoken BNC2014*. Routledge, New York.

Daan Broeder, Thierry Declerck, Erhard Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari, and Peter Wittenburg. 2008. Foundation of a Component-based Flexible Registry for Language Resources and Technology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1433–1436, Marrakech, Morocco.

Piotr Bánski, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Piotr P zik, Carsten Schnober, and Andreas Witt. 2013. KorAP: the new corpus analysis platform at IDS Mannheim. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human language technology challenges for computer science and linguistics. 6th language & technology conference*, pages 586–587. Uniwersytet im. Adama Mickiewiczaw Poznaniu, Poznań.

Andrew Chesterman. 1998. *Contrastive Functional Analysis*. Number 47 in Pragmatics & Beyond. Benjamins, Amsterdam.

Ruxandra Cosma, Dan Cristea, Marc Kupietz, Dan Tufiş, and Andreas Witt. 2016. DRuKoLA – towards contrastive German-Romanian research

based on comparable corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, pages 28–32, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals*, 60(2):95–99.

Mark Davies. 2011. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, 25:447–465.

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP Architecture - Diving in the Deep Sea of Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3586–3591, Portorož, Slovenia. European Language Resources Association (ELRA).

Stefan Evert. 2006. How Random is a Corpus? The Library Metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54(2).

William Frawley. 1992. *Linguistic semantics*. Lawrence Erlbaum Associates, Hillsdale.

Silviane Granger and Marie-Aude Lefer, editors. 2022. *Extending the Scope of Corpus-Based Translation Studies*. Bloomsbury Advances in Translation. Bloomsbury, London, UK.

Sylviane Granger. 2010. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Contemporary Foreign Language Studies*, 10(2):14–21.

Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson. 2003. *Corpus-based approaches to contrastive linguistics and translation studies*, volume 20. Rodopi, Amsterdam & Atlanta.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proceedings of the 5th workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, pages 35–42, Stroudsburg. ACL.

Carl James. 1980. *Constrative Analysis*. Longman, London.

Daniel Janus and Adam Przepiórkowski. 2007. Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 85–88, Prague, Czech Republic. Association for Computational Linguistics.

John Kirk and Anna Čermáková. 2017. From ICE to ICC: The new International Comparable Corpus. In Piotr Bański, Marc Kupietz, Harald Lüngen, Paul Rayson, Hanno Biber, Evelyn Breiteneder, Simon Clematide, John Mariani, Mark Stevenson, and Theresa Sick, editors, *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*, pages 7 – 12. IDS, Mannheim.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Alexander Koplenig. 2017. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15(2):321–346.

Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. 2016. Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*, 29:ecw029.

Marc Kupietz. 2015. Constructing a Corpus. In Philip Durkin, editor, *The Oxford Handbook of Lexicography*, pages 62–75. Oxford University Press.

Marc Kupietz, Adrien Barbaresi, Anna Čermáková, Małgorzata Czachor, Nils Diewald, Jarle Ebeling, Rafał L. Górski, Eliza Margaretha, John Kirk, Michal Křen, Harald Lüngen, Signe Oksefjell Ebeling, Mícheál Ó Meachair, Ines Pisetta, Elaine Uí Dhonnchadha, Friedemann Vogel, Rebecca Wilm, Jiajin Xu, and Rameela Yaddehige. 2023. News from the International Comparable Corpus. First launch of ICC written. In *Book of Abstracts of the 10th International Contrastive Linguistics Conference (ICLC-10)*, pages 45–48, Mannheim, Germany. IDS-Verlag; Leibniz-Institut für Deutsche Sprache (IDS).

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta. European Language Resources Association (ELRA).

Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2020a. RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo

via KorAP. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, pages 7015–7021, Marseille, France. European Language Resources Association.

Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2022. Building paths to corpus data: A multi-level least effort and maximum return approach. In Darja Fišer and Andreas Witt, editors, *CLARIN. The Infrastructure for Language Resources.*, pages 163–189. deGruyter, Berlin. Section: number x.

Marc Kupietz, Nils Diewald, Beata Trawiński, Ruxandra Cosma, Dan Cristea, Dan Tufiş, Tamás Váradi, and Angelika Wöllstein. 2020b. Recent developments in the European Reference Corpus EuReCo. *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*, pages 257–273.

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German reference corpus DeReKo: New developments – new opportunities. In *Proceedings of the Eleventh International Conference on language resources and evaluation (LREC '18)*, pages 4353–4360, Miyazaki, Japan. ELRA.

Marc Kupietz and Beata Trawiński. 2022. Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo. In Laura Auteri, Natascia Barrale, Arianna Di Bella, and Sabine Hoffmann, editors, *Wege der Germanistik in transkultureller Perspektive. Akten des XIV. Kongresses der Internationalen Vereinigung für Germanistik (IVG) (Bd. 6)*, Jahrbuch für Internationale Germanistik - Beihefte - 6, pages 417–439. Peter Lang, Bern.

Marc Kupietz, Andreas Witt, Piotr Bański, Dan Tufiş, Dan Cristea, and Tamás Váradi. 2017. EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In *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*, pages 15–19, Mannheim. Institut für Deutsche Sprache.

Michal Křen. 2020. Czech National Corpus in 2020: Recent Developments and Future Outlook. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 52–57, Marseille, France. European Language Ressources Association.

Sara Laviosa. 1998. Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, 43(4):557–570.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Anthony M. McEnery and Richard Zhonghua Xiao. 2007. Parallel and comparable corpora: What is Happening? In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: The Linguist and the Translator*. Multilingual Matters, Clevedon, UK.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, pages 1719–1723, Reykjavik, Iceland. European Language Resources Association (ELRA).

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Alexandr Rosen, Martin Vavřín, and Adrian J. Zasina. 2019. *The InterCorp Corpus – Czech1, 12. Version*. Institute of the Czech National Corpus/Charles University, Prague.

Pavel Rychlý. 2007. Manatee / Bonito - A modular corpus manager. In Petr Sojka and Aleš Horák, editors, *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70. Masaryk University, Brno.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004. A Language Resources Infrastructure for Bulgarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Lora Spassova. 2023. *Integrating a Large Bulgarian Corpus into the European Reference Corpus EuReCo*. Bachelor Thesis, Heinrich-Heine University Düsseldorf.

Janusz Taborek. 2020. Kookkurrenz und syntagmatische Muster der Funktionsverbgefüge aus kontrastiver deutsch-polnischer Sicht am Beispiel in Not geraten. In Sabine Knop and Manon Hermann, editors, *Funktionsverbgefüge im Fokus:*

Theoretische, didaktische und kontrastive Perspektiven, pages 211–234. De Gruyter, Berlin.

TEI Consortium. 2024. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

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

Wolfgang Teubert. 1998. Korpus und Neologie. In Wolfgang Teubert, editor, *Neologie und Korpus*, number 11 in Studien zur deutschen Sprache, pages 129–170. Narr, Tübingen.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. ELRA.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus – parallel & free. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, pages 1183–1186, Lisbon, Portugal. ELRA.

Beata Trawiński and Marc Kupietz. 2021. Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In Henning Lobin, Andreas Witt, and Angelika Wöllstein, editors, *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch*, number 2020 in Jahrbuch / Leibniz-Institut für Deutsche Sprache (IDS), pages 209–234. de Gruyter, Berlin, Germany.

Beata Trawiński, Marc Kupietz, Kristel Proost, and Jörg Zinken, editors. 2023. *10th International Contrastive Linguistics Conference (ICLC). Book of abstracts*. IDS, Mannheim, Germany.

Thorsten Trippel. 2013. Minutes to the Workshop on Federated Content Search. Technical report, University of Copenhagen, Copenhagen.

Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei. 2019. Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. *On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*, 64(3). Place: Bucharest, Romania Publisher: Editura Academiei Române.

Martin Volk, Anne Göhring, Annette Rios, Torsten Marek, and Yvonne Samuelsson. 2015. *SMULTRON (4. Version) — The Stockholm MULtilingual parallel TReebank*. Institute of Computational Linguistics, University of Zurich, Zurich.

Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 385–389, Las Palmas, Spain. European Language Resources Association (ELRA).

Marcin Woliński. 2014. Morfeusz Reloaded. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1106–1111, Reykjavik, Iceland. European Language Resources Association (ELRA).

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Anna Čermáková, Jarmo Jantunen, Tommi Jauhiainen, John Kirk, Michal Křen, Marc Kupietz, and Elaine Uí Dhonnchadha. 2021. The International Comparable Corpus: Challenges in building multilingual spoken and written comparable corpora. *Research in Corpus Linguistics: Special issue "Challenges of combining structured and unstructured data in corpus development"*, 9(1):89 – 103.