News from EuReCo:
Annotations, Applications, and LLM Assistance

Keywords: European Languages; Light Verb Constructions; Comparable Corpora; Tools; LLMs

The field of contrastive corpus linguistics is inherently more resource-intensive than single-language studies due to the necessity of at least two corpora that are not only sufficiently representative with respect to the research question and intended language domain but also sufficiently similar. While parallel or translation corpora exist for many languages and domains (cf., Čermák/Rosen 2012) and meet the similarity requirement, their linguistic utility is often affected by translation effects, such as shining-through, over-normalization, and simplification (e.g., Teich, 2003; Granger et al., 2003). Comparable corpora present a more effective alternative for capturing authentic cross-linguistic patterns; however, locating or creating such corpora for specific language constellations and domains can be highly costly and labor-intensive.

The European Reference Corpus (EuReCo) open initiative (Kupietz et al. 2020) provides a sustainable solution to this challenge by leveraging existing large corpora, virtually integrating them, and enabling users to define domain- and question-specific virtual sub-corpora based on metadata properties. This approach eliminates the need to build new corpora from scratch while ensuring economic feasibility. EuReCo's strategy involves dynamically joining national and reference corpora, allowing each corpus to remain physically decentralized yet interoperable through infrastructural means. To ensure the feasibility of developing and maintaining the software that implements this infrastructure, EuReCo adopts a more centralized approach. Rather than specifying interfaces and protocols for interoperable searches across the network, EuReCo offers a prototype open-source implementation, KorAP (Bański et al. 2012, Diewald et al. 2016), which can be installed at the locations of the corpora to provide (in most cases additional) access to the corpus data (Kupietz et al. 2024). This method is more efficient and guarantees that new features are available to all users in a timely manner without multiplying costs across participating sites or software systems.

The contribution presents fundamental approaches of EuReCo in addressing these challenges and reports on ongoing research related to methodological feasibility issues, particularly from a user perspective. This includes enhancing metadata on topic domains through multilingual text classifiers for constructing more precise virtual sub-corpora. Additionally, it addresses query-level mapping of Universal Part-of-Speech (UPOS) annotations to facilitate cross-linguistic comparisons without necessitating re-annotation of entire corpora, as well as LLM-supported queries and scripted access for traceable and reproducible results by incorporating LLMs through variations of Retrieval-Augmented Generation (RAG) and specialized interfaces like Model Concept Protocol (MCP) to streamline complex data analysis tasks. Furthermore, we will report on ongoing research concerning light-verb constructions (drawing on Bański et al., 2023), aspect(uality), and negation based on available corpora from Bulgarian, Finnish, German, Hungarian, Polish, and Romanian.

References

Bański, Piotr/Fischer, Peter M./Frick, Elena/Ketzan, Erik/Kupietz, Marc/Schnober, Carsten/Schonefeld, Oliver/Witt, Andreas (2012): The New IDS Corpus Analysis Platform: Challenges and Prospects. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). *LREC 2012*, Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2905–2911. http://www.lrec-conf.org/proceedings/lrec2012/pdf/789_Paper.pdf.

Bański, Piotr/Diewald, Nils/Kupietz, Marc/Trawiński, Beata (2023): Applying the newly extended European reference corpus EuReCo. Pilot studies of light-verb constructions in German, Romanian, Hungarian and Polish. in Trawiński, Beata/Kupietz, Marc/Proost, Kristel/Zinken, Jörg (Hrsg.): Book of Abstracts of the 10th International Contrastive Linguistics Conference (ICLC-10), 18-21 July, 2023, Mannheim, Germany. Mannheim: IDS-Verlag, S. 274–276. https://doi.org/10.14618/f8rt-m155.

Čermák, František/Rosen, Alexandr (2012): The case of InterCorp, a multilingual parallel corpus. In: International Journal of Corpus Linguistics, 17(3), pp. 411–427. https://doi.org/10.1075/ijcl.17.3.05cer.

Diewald, Nils/Hanl, Michael/Margaretha, Eliza/Bingel, Joachim/Kupietz, Marc/Bański, Piotr/Witt, Andreas (2016): KorAP architecture - Diving in the Deep Sea of Corpus Data. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, pp. 3586–3591.

Granger, Sylviane/Lerot, Jacques/Petch-Tyson, Stephanie (2003): Corpus-based approaches to contrastive linguistics and translation studies. Amsterdam & Atlanta: Rodopi.

Kupietz, Marc/Bański, Piotr/Diewald, Nils/Trawiński, Beata/Witt, Andreas (2024): EuReCo: Not building and yet using federated comparable corpora for cross-linguistic research. In: Zweigenbaum, Pierre/Rapp, Reinhard/Sharoff, Serge (eds.): Proceedings of the BUCC 2024: The 17th workshop on building and using comparable corpora. Torino, Italia: ELRA Language Resource Association, pp. 94–103. https://aclanthology.org/2024.bucc-1.10.pdf.

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: 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, pp. 257–273.

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