

News from the International Comparable Corpus: First launch of ICC written

The International Comparable Corpus (ICC) (Kirk/Čermáková 2017; Čermáková et al. 2021) is an open initiative which aims to improve the empirical basis for contrastive linguistics by compiling comparable corpora for many languages and making them as freely available as possible as well as providing tools with which they can easily be queried and analysed. In this contribution we present the first release of written language parts of the ICC which includes corpora for Chinese, Czech, English, German, Irish, and Norwegian. Each of the released corpora contains 400k words distributed over 14 different text categories according to the ICC specifications. Our poster covers the design basics of the ICC, its TEI encoding, a demonstration of using the ICC via different query tools, and an outlook on future plans.

Similar to the European Reference Corpus EuReCo (Kupietz et al. 2020), ICC follows the approach of reusing existing linguistic resources wherever possible in order to cover as many languages as possible with realistic effort in as short a time as possible. In contrast to EuReCo, however, comparable corpus pairs are not defined dynamically in the usage phase, but the compositions of the corpora are fixed in the ICC design. The approaches are thus complementary in this respect. The design principles and composition of the ICC are based on those of the International Corpus of English (ICE) (Greenbaum 1996), with the deviation that the ICC includes the additional text category blog post and excludes spoken legal texts (see Čermáková et al. 2021 for details). ICC's fixed-design approach has the advantage that all single-language corpora in the ICC have the same composition with respect to the selected text types and that this guarantees that the selected broad spectrum of potential influencing variables for linguistic variation is always represented. The disadvantage, however, is that this can only be achieved for quite small corpora and that the generalisability of comparative findings based on the ICC corpora will often need to be checked on larger monolingual corpora or translation corpora (Čermáková/Ebeling/Ebeling forthcoming). Arguing that such issues with comparability and representativeness are inevitable, in one way or the other, and need to be dealt with, our poster will discuss and exemplify the text selections in more detail.

ICC's original aim was to make all corpora available for download. However, this goal turned out to be unfeasible, as it was often not possible to obtain licences for individual texts with reasonable effort. In addition, the copyright exceptions are too different to find a uniform solution for sharing full texts. In order to come as close as possible to our original goal, we have thus decided to make the ICC accessible at least via several corpus platforms and on several access levels (Kupietz/Diewald/Margaretha 2022), requiring users only to electronically sign an end-user licence agreement that provides for exclusively academic, non-commercial use. Our poster will demonstrate ICC access via the corpus platforms KonText (Machálek 2020), KorAP (Diewald et al. 2016), and Korp (Borin/Forsberg/Roxendal 2012) showing exemplary comparative analyses of light verb constructions in selected ICC-corpora using native and Universal Dependency annotations (Nivre et al. 2020).

In the final part of the poster, we discuss the plans of future ICC extensions, intensifying the relations with the EuReCo, and the roadmap for completing the spoken language parts of ICC.

References

- Borin, L./Forsberg, M./Roxendal J. (2012): Korp – the corpus infrastructure of Språkbanken. Proceedings of LREC 2012. Istanbul: ELRA, pp 474–478.
- Čermáková, A./Ebeling, J./Ebeling Oksefjell, S. (Forthcoming 2023). ‘Be’ verbs in a contrastive perspective: The case of *být*, *be* and *være*. *Nordic Journal of English Studies*.
- Čermáková, A./Jantunen, J./Jauhiainen, T./Kirk, J./Křen, M./Kupietz, M./Dhonnchadha, E.U. (2021): The International Comparable Corpus: Challenges in building multilingual spoken and written comparable corpora. In: Research in Corpus Linguistics: Special issue ‘Challenges of combining structured and unstructured data in corpus development’. Edited by T. Säily/J. Tyrkkö, 9 (1), pp. 89–103. <http://doi.org/10.32714/ricl.09.01.06>.
- 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: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). Portorož, Slovenia, pp. 3586–3591. <https://www.aclweb.org/anthology/L16-1569>.
- Greenbaum, S. (ed.) (1996): Comparing English Worldwide: The International Corpus of English. Oxford.
- Kirk, J./Čermáková, A. (2017): From ICE to ICC: The new International Comparable Corpus. In: Bański, P./Kupietz, M./Lüngen, H./Rayson, P./Biber, H./Breiteneder, E./Clematide, S./Mariani, J./Stevenson, M./Sick, T. (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017, pp. 7–12.
- Kupietz, M./Diewald, N./Margaretha, E. (2022): Building paths to corpus data: A multi-level least effort and maximum return approach. In: Fišer, D./Witt, A. (eds.): CLARIN. The Infrastructure for Language Resources. Berlin. <https://doi.org/10.1515/9783110767377-007>.
- Kupietz, M./Diewald, N./Trawiński, B./Cosma, R./Cristea, D./Tufiş, D./Váradi, T./Wöllstein, A. (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.
- Machálek, T. (2020): KonText: Advanced and Flexible Corpus Query Interface. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France, pp. 7003–7008. <https://www.aclweb.org/anthology/2020.lrec-1.865>.
- Nivre, J./de Marneffe, M.-C./Ginter, F./Hajič, J./Manning, C.D./Pyysalo, S./Schuster, S./Tyers, F./Zeman, D. (2020): Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France, pp. 4034–4043. <https://www.aclweb.org/anthology/2020.lrec-1.497>.
- Teich, E. (2003): Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts. Berlin.