Marc Kupietz / Adrien Barbaresi / Anna Cermakova /
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 / Rameela Yaddehige

News from the International Comparable Corpus

First launch of ICC written

Keywords: comparable corpora; international comparable corpus; contrastive linguistics; corpus linguistics; linguistic research software

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 (partly), 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 Oksefjell forthcoming). Arguing that such issues with comparability and representa-

tiveness 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 query and analysis platform KorAP[1] (Diewald et al. 2016) showing exemplary comparative analyses of light verb constructions in selected ICC-corpora using Universal Dependency annotations (Nivre et al. 2020) provided by UDPipe 2.0 (Straka 2018). As further platforms we plan to add KonText (Machálek 2020) and Korp (Borin/Forsberg/Roxendal 2012).

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, Lars/Forsberg, Markus/Roxendal, Johan (2012): Korp — the corpus infrastructure of Språkbanken. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA), pp. 474–478. http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf.

Čermáková, Anna/Ebeling, Jarle/Ebeling Oksefjell, Signe (forthcoming): 'Be' verbs in a contrastive perspective: The case of být, be and være. In: Nordic Journal of English Studies.

Čermáková, Anna/Jantunen, Jarmo/Jauhiainen, Tommi/Kirk, John/Křen, Michal/Kupietz, Marc/Uí Dhonnchadha, Elaine (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 Tanja Säily/Jukka Tyrkkö, 9(1), pp. 89–103. https://doi.org/10.32714/ricl.09.01.06.

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'16). Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3586–3591. https://www.aclweb.org/anthology/L16-1569.

Greenbaum, Sidney (ed.) (1996): Comparing English Worldwide: The International Corpus of English. Oxford: Clarendon Press.

Kirk, John/Čermáková, Anna (2017): From ICE to ICC: The new International Comparable Corpus. In: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/Stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017. IDS, pp. 7–12. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62490.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2022): Building paths to corpus data: A multi-level least effort and maximum return approach. In: Fišer, Darja/Witt, Andreas (eds.): CLARIN. The Infrastructure for Language Resources. Berlin: deGruyter. https://doi.org/10.1515/9783110767377-007.

[1] You can perform your own ICC queries and analysis via https://korap.ids-mannheim.de/instance/icc/

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.

Machálek, Tomáš (2020): KonText: Advanced and Flexible Corpus Query Interface. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865.

Nivre, Joakim/de Marneffe, Marie-Catherine/Ginter, Filip/Hajič, Jan/Manning, Christopher D./Pyysalo, Sampo/Schuster, Sebastian/Tyers, Francis/Zeman, Daniel (2020): Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 4034–4043. https://www.aclweb.org/anthology/2020.lrec-1.497.

Straka, Milan (2018): UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium: Association for Computational Linguistics, pp. 197–207. https://doi.org/10.18653/v1/K18-2020.

Contact information

Marc Kupietz
Leibniz-Institut für Deutsche Sprache
kupietz@ids-mannheim.de

Adrien Barbaresi
Berlin Brandenburg Academy of Sciences
barbaresi@bbaw.de

Anna Cermakova
Charles University
anna.cermakova@ff.cuni.cz

Małgorzata Czachor
Institute of Polish Language, Polish Academy of Sciences
malgorzata.czachor@ijp.pan.pl

Nils Diewald
Leibniz-Institut für Deutsche Sprache
diewald@ids-mannheim.de

Jarle Ebeling
University of Oslo
jarle.ebeling@usit.uio.no

Signe Oksefjell Ebeling
University of Oslo
s.o.ebeling@ilos.uio.no

Rafał L. Górski
Institute of Polish Language, Polish Academy of Sciences
rafal.gorski@ijp.pan.pl

John Kirk
University of Vienna
john.kirk@univie.ac.at

Michal Křen
Charles University
michal.kren@ff.cuni.cz

Harald Lüngen
Leibniz-Institut für Deutsche Sprache
luengen@ids-mannheim.de

Eliza Margaretha
Leibniz-Institut für Deutsche Sprache
margaretha@ids-mannheim.de

Mícheál Ó Meachair
Dublin City University
micheal.omeachair@dcu.ie

Ines Pisetta
Leibniz-Institut für Deutsche Sprache
pisetta@swhk.ids-mannheim.de

Elaine Uí Dhonnchadha
Trinity College Dublin
uidhonne@tcd.ie

Friedemann Vogel
University of Siegen
friedemann.vogel@uni-siegen.de

Rebecca Wilm
Leibniz-Institut für Deutsche Sprache
wilm@ids-mannheim.de

Jiajin Xu
The National Research Centre for Foreign Language Education, Beijing Foreign Studies University, China
xujiajin@bfsu.edu.cn

Rameela Yaddehige
Leibniz-Institut für Deutsche Sprache
yaddehige@ids-mannheim.de