News from the International Comparable Corpus

First launch of ICC written

Marc Kupietz[1], Adrien Barbaresi[2], Anna Čermáková[3], Małgorzata Czachor[4], Nils Diewald[1], Jarle Ebeling[5], Rafał L. Górski[4], John Kirk[6], Michal Křen[3], Harald Lüngen[1], Eliza Margaretha[1], Signe Oksefjell Ebeling[5], Mícheál Ó Meachair[7], Ines Pisetta[1], Elaine Uí Dhonnchadha[8], Friedemann Vogel[9], Rebecca Wilm[1], Jiajin Xu[10], Rameela Yaddehige[1]

[1] IDS Mannheim; [2] BBAW Berlin; [3] Charles University; [4] Polish Academy of Sciences; [5] University of Oslo; [6] University of Vienna; [7] Dublin City University; [8] Trinity College Dublin; [9] University of Siegen; [10] Beijing Foreign Studies University

ICC AIMS & CHARACTERISTICS

- make available comparable corpora of many languages for contrastive linguistic research (Kirk/Čermáková 2017)
- mostly based on existing corpora
- small corpora with 1M words each (400K written)
- pre-defined “balanced” composition
  - inspired by the International Corpus of English (Greenbaum 1996)

CURRENT LAUNCH OF ICC WRITTEN

- written parts for Chinese, Czech, English (mostly), German, Irish (partly), Norwegian publicly available
  - partially including UDPipe 2.0 annotations (Straka 2018)
  - usable via CWB or KorAP (Diewald et al. 2016), see QR Code

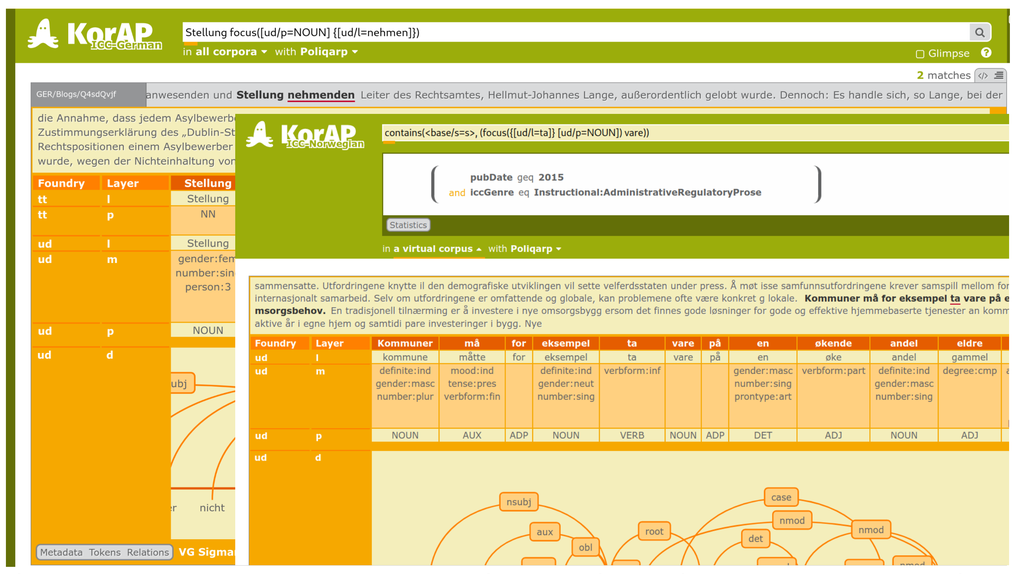

Figure 1: KorAP UI for ICC-GER and ICC-NOR, showing annotation queries and layers, as well as a virtual corpus definition, based on ICC genre and publication date metadata.

Composition of the ICC parts

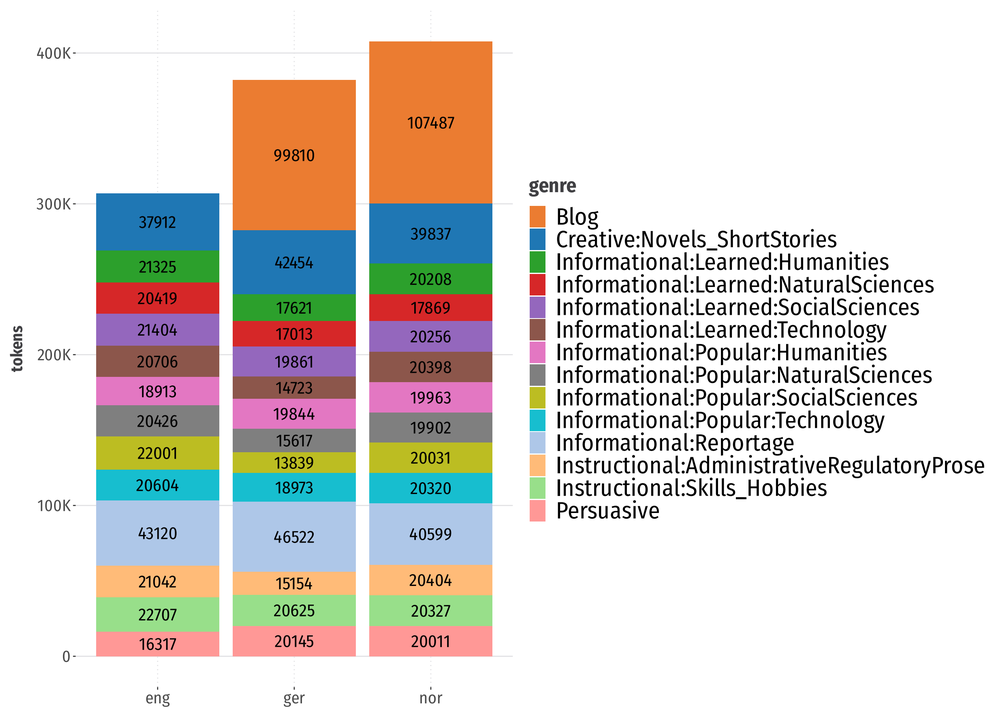

Figure 2: Actual composition of selected ICC parts with respect to ICC domain. (For the other ICC parts, the ICC genre metadatum was not yet accessible via the API at the editorial deadline.)

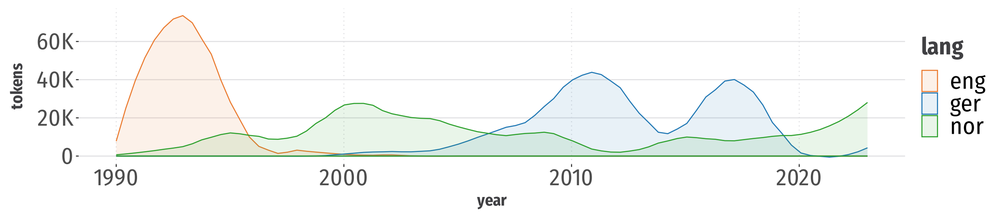

Figure 3: Composition of the selected ICC parts with respect to year of publication.

PILOT STUDY

- identification of light verb constructions (LVC) with *take* in English, and corresponding lemmas in German and Norwegian
  - to explore the limitations imposed by the small corpus sizes
  - using RKorAPClient (Kupietz/Diewald/Margaretha 2020) to access the corpora and get reproducible results for the analyses

```
library(RKorAPClient)
new("KorAPConnection",
    KorAPUrl = "https://korap.ids-mannheim.de/instance/icc/eng",
    accessToken = Sys.getenv("KORAP_ICC_TOKEN_eng")) %>%
collocationAnalysis(
    "focus({[ud/l=take]} [ud/p=NOUN])",
    leftContextSize = 0,
    rightContextSize = 1,
    minOccur = 2,
    addExamples = T)
```

	Collocate	Example	logDice	pmi	ll
1	place	the changes taking **place** in	12.40	9.59	537.15
2	part	full participation (taking **part**) of	10.57	7.63	127.55
3	care	farmers to take **care** when	10.22	8.08	60.90
4	account	these services take **account** of	9.86	8.04	41.85
5	precautions	duties to take **precautions** against	9.81	9.27	38.94
6	advantage	society to take **advantage** of	9.69	8.49	34.82
7	action	around before taking **action**?	9.50	7.71	30.84
8	pride	Protestants to take **pride** in	9.38	9.34	28.09
9	control	choose to take **control** of	8.55	6.01	16.03
10	time	something which takes **time** and	7.68	4.41	14.86

Figure 4: R code for, and results of a co-occurrence analysis of *take* + NOUN in ICC-ENG, using the RKorAPClient package.

Results

- for English the query for *take* + NOUN (as direct right neighbour) yields 10 different pairs with a minimum frequency of 2 (see Figure 4)
  - based on English Wikipedia (2015 snapshot, see Margaretha/Lüngen 2014) the query yields 139 pairs (log-dice-threshold: 2.0) with 44 false positives
  - the true positive ratio of discovered take-LVCs between ICC and Wikipedia is 10:95
- for ICC German with DeReKo as background corpus, the ratio of discovered true LVCs with ›nehmen‹ (=take) is 10:89
- in both cases, not much more than 10% of LVCs could be discovered

SUMMARY & OUTLOOK

- we have made comparable corpora of 4+ languages available, readily usable for contrastive research
- however, even for fairly frequent phenomena, the results on the small corpora should be treated with caution
  - typically, they need to be verified on larger monolingual corpora
  - this also and especially concerns recall
- nevertheless ICC can serve as a useful basis for contrastive studies
  - with a uniform UI and API that facilitate query and analysis
- in addition, ICC also serves as a crystallisation point
  - for more ICC corpora and spoken parts to come
  - for larger corpora and complementary approaches, such as EuReCo

REFERENCES

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 / Paris: ELRA, pp. 3586–3591. https://aclanthology.org/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 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 (2020): RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP. in Proceedings of the 12th Language Resources and Evaluation Conference. Marseille / Paris: ELRA, pp. 7015–7021. https://aclanthology.org/2020.lrec-1.867/.

Margaretha, Eliza/Lüngen, Harald (2014): Building linguistic corpora from wikipedia articles and discussions. in Journal of Language Technology and Computational Linguistics. Special issue on building and annotating corpora of computer-mediated communication. Issues and challenges at the interface between computational and corpus linguistics. Edited by Michael Beißwenger/Angelika Storrer/Nelleke Oostdijk/Henk van den Heuvel, 29(2), pp. 59–82. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-33306.

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:
Dr. Marc Kupietz
Corpus Linguistics PA
Leibniz-Institut für Deutsche Sprache
PO Box 10 16 21
68016 Mannheim
Germany

https://www.ids-mannheim.de/digspra/kl
icc@ids-manneim.de

Street Address:
Leibniz-Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
Germany

Phone: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de