IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Assembling EuReCo for Contrastive Research: The Polish Piece

Piotr Bański, Nils Diewald, Marc Kupietz and Beata Trawiński

EuReCo is a federated virtual corpus under development since 2013, enabling large user-definable comparable corpora across various languages upon an infrastructure supporting contrastive research. The core of the infrastructure is KorAP, a scalable open-source corpus platform. So far, EuReCo comprises the German Reference Corpus DeReKo, the Reference Corpus of Contemporary Romanian Language and the Hungarian National Corpus.

EURECO APPROACH TO COMPARABLE CORPORA

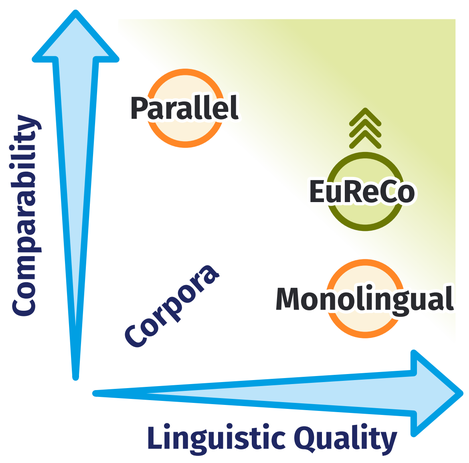
- **open initiative** founded in 2013
- **combine existing national and reference corpora**
  - maintained by sustainable institutions
  - instead of building new ones
- provide predefined virtual (comparable) corpora **AND**
- enable users to define **specific virtual comparable corpus pairs** dynamically

BECAUSE ...

- corpora with reasonable size and diversity **cannot be perfectly comparable in general**
  - always at least one criterion against comparability
  - whether unequal distribution with regard to such a variable is relevant **depends on the specific research question**
- also: single language corpora cannot be generally representative
  - population is not generally definable
  - whether a corpus is sufficiently representative **depends on the research question and the language domain**

BENEFITS OF THE EURECO APPROACH

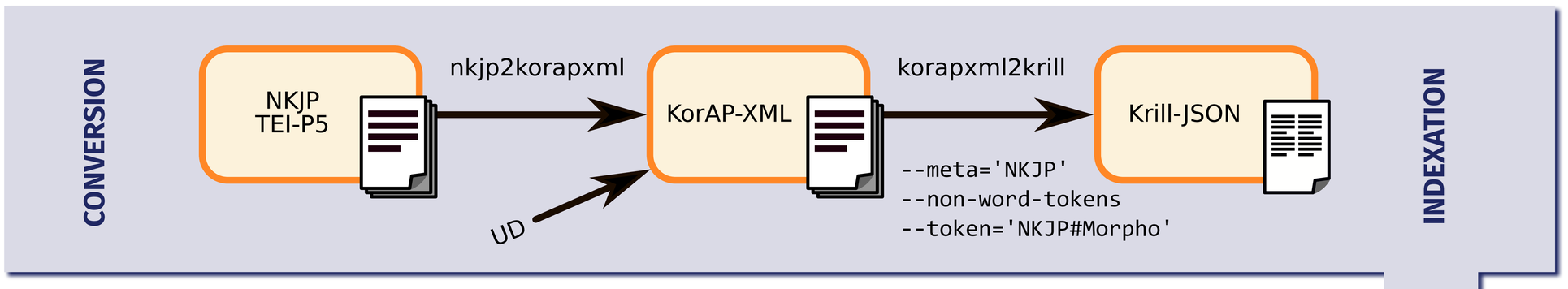
- more **economical**, **scalable** and **sustainable**
  - especially since one can also benefit from ongoing and future extensions and improvements of these corpora
  - high linguistic quality and sufficient size to be expected



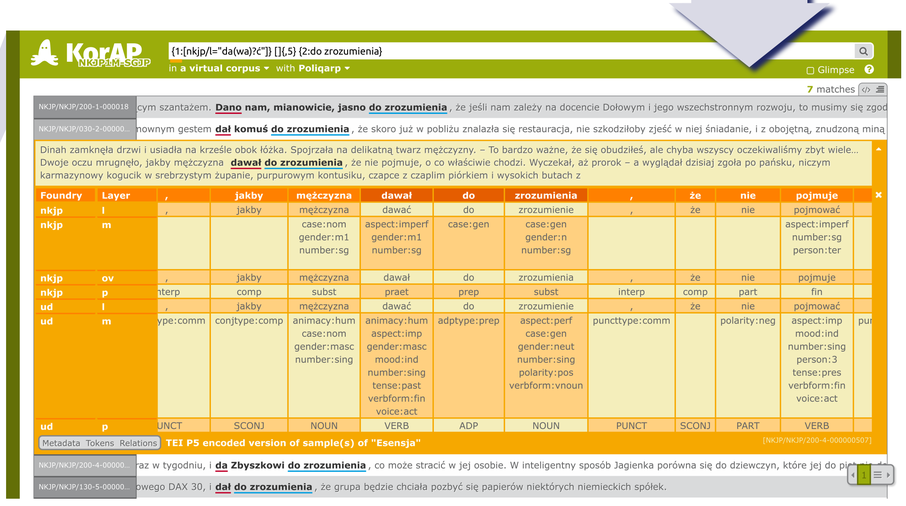
APPROACHING COMPARABILITY

- draw sub-corpora from monolingual corpora
  - so that they have similar token distributions with respect to **metadata variables** like topic area, text type, publication date etc.
- Iterative refinement for gradual approximation to sufficient comparability concerning specific research questions

NKJP-1M-SGJP IN KORAP



- Conversion for the NKJP corpus was conducted in multiple steps by extending the existing KorAP ingestion pipeline:
  1. NKJP format was transformed to KorAP-XML using **dedicated XSLT scripts**
  2. KorAP-XML data was enriched using **UDPipe 2** for Polish
  3. KorAP-XML data was further transformed using korapxml2krill with **dedicated conversion rules** for metadata and annotations
- To improve access to the NKJP annotation scheme, the annotation assistant for KorAP was extended to support the tagsets for part-of-speech and morphological information



ANALYSIS WITH R

```
01 library(RKorAPClient)
02
03 new(
04   "KorAPConnection",
05   KorAPUrl = "https://korap.ids-mannheim.de/instance/nkjp1m-sgjp"
06 ) %>%
07 collocationAnalysis(
08   'focus({[nkjp/l="da(wa)?ć"] []{,5} do} [nkjp/p=subst])',
09   leftContextSize = 0,
10   rightContextSize = 1, # relative to { ... } in focus()
11 )
```

LVC	logDice	pmi	ll
da(wa)?ć ... do zrozumienia	11.84246	13.05918	107.977

FURTHER INFORMATION

Bański, P., Fischer, P. M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O., Witt, A., 2012. **The New IDS Corpus Analysis Platform: Challenges and Prospects**, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2905–2911.

Barbu Mititelu, V., Tufiş, D., Irimia, E., 2018. **The Reference Corpus of the Contemporary Romanian Language (CoRoLa)**, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

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). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3586–3591.

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**. Transl. Comp. Lang. Corpus-Based Insights Sel. Proc. Fifth Using Corpora Contrastive Transl. Stud. Conf. Louvain--Neuve Press. Univ. Louvain, Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference, pp. 257–273.

Kupietz, M., Lüngen, H., Kamocki, P., Witt, A., 2018. **The German Reference Corpus DeReKo: New Developments – New Opportunities**, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

Kupietz, M. and Trawiński, B., 2022. **Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo**, in: Janusz Taborek, Henning Lobin, Fabio Mollica (Eds.): Kontrastive Korpuslinguistik. Akten des XIII. Internationalen Germanistenkongresses Shanghai 2015: Germanistik zwischen Tradition und Innovation. Peter Lang Verlag, pp. 417-439.

Przepiórkowski, A., Górski, R.L., Łaziński, M., Pęzik, P., 2010. **Recent Developments in the National Corpus of Polish**, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta.

Váradi, T., 2002. **The Hungarian National Corpus**, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain.

https://www.ids-mannheim.de/digspra/kl/projekte/eureco

Contact:
Nils Diewald
Leibniz Institute for the German Language
Dpt. for Digital Linguistics
PO Box 10 16 21
D-68016 Mannheim
Germany

Phone: +49 621 1581-450
diewald@ids-mannheim.de

Street Address:
Leibniz Institute for the German Language
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

Leibniz Gemeinschaft