

Building and querying Wikipedia discussion corpora using KorAP

Eliza Margaretha, Harald Lungen, Nils Diewald, Marc Kupietz, Rameela Yaddehige

Leibniz-Institut für Deutsche Sprache
Mannheim

(margaretha|luengen|diewald|kupietz|yaddehige)@ids-mannheim.de

Abstract

We introduce the new German Wikipedia talk page corpus with 1.14 billion tokens and multiple linguistic annotation layers, available via the corpus analysis platform KorAP.

Keywords: Wikipedia, talk pages, wikitext, CMC, corpus construction, KorAP

The 349 language versions (as of August 2024) of the online encyclopedia Wikipedia are consulted by hundreds of thousands of users daily and serve as the foundation of major LLMs. They are a huge collaborative effort with thousands of authors who contribute on a voluntary basis. Besides the encyclopaedic articles, Wikipedia contains talk pages (a.k.a. discussions), i.e. a namespace where the authors discuss and negotiate the composition of articles (article talk) or discuss more freely (user talk). Unlike the articles, talk pages are organised in a dialogue structure with postings and threads, hence they form a very large and interesting linguistic archive of CMC (Lungen and Kupietz, 2017; Ho-Dac, 2024).

The poster presents the latest article and user discussion corpora of the German Wikipedia we built in 2024, comprising 1.14 billion tokens in 1.5 million documents. The corpus is accessible through KorAP (Bański et al., 2012), a web-based corpus analysis platform. In addition to a graphical user interface, KorAP provides API web-services that allow users to access corpora using client applications such as RKorAP-Client (Kupietz et al., 2020) to extract and visualize the results in these applications.

Enhanced Wikipedia corpus builder

KorAP takes I5, the TEI customisation for the German Reference Corpus DEREKO (Lungen and Sperberg-McQueen, 2012) as import format. The Wikipedia sources come as database dumps from <https://dumps.wikimedia.org/> and contain all pages in the wikitext format. Figure 1 illustrates our conversion pipeline comprising five modules: pre-processing, parsing, XML rendering, transformation, and post-processing. In pre-processing, HTML tags are escaped to retain structure, missing tags are corrected using TagSoup¹, and posting segmentation is applied. Subsequently, the wikitext is parsed using Sweble² into an abstract syntax tree (AST), which is eventually rendered as XML, producing a WikiXML corpus. An

¹<http://vrici.lojban.org/~cowan/tagsoup/>

²Sweble (Dohrn and Riehle, 2011) is a parser for wikitext documents. It has been updated recently for compatibility with Java 17, but is no longer under active development. <https://github.com/sweble>

I5 Wikipedia corpus is finally produced through XSL transformation followed by post-processing to handle categories and cross-language links.

Querying emojis and emoticons in KorAP

Talk pages exhibit orthographic and lexical features typical of CMC including emoticons and emojis. Plain text emoticons (e.g. :-)) are searchable in KorAP. KorAP also supports searching emojis by using Unicode (see Figure 2), however, only a small number of emojis are encoded in Unicode (e.g. 😊) in wikitext. Most emojis are encoded as templates (e.g. `{S}`), which requires special processing in tokenization to enable their searchability, and a normalization to Unicode to improve visualization.

The KorAP instance that contains the latest wiki corpora is located at <https://korap.ids-mannheim.de/instance/wiki>.³

Extracting and visualising results using RKorAPClient

The poster will feature analyses that compare linguistic properties of the article talk pages, the user talk pages, and a press-subcorpus from the German Reference Corpus DEREKO (Kupietz et al., 2010).

References

- Bański, P., Fischer, P. M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O., and 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)*, pages 2905–2911. European Language Resources Association (ELRA).
- Dohrn, H. and Riehle, D. (2011). Design and implementation of the Sweble Wikitext parser: unlocking the structured data of Wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11*, pages 72–82. ACM.
- Ho-Dac, L.-M. (2024). Building a comparable corpus of online discussions on Wikipedia. In *Investigating*

³In addition, the corpora are available for download at <https://www.ids-mannheim.de/en/digspra/pb-s1/projects/corpus-development/verfuegbarkeit/>.

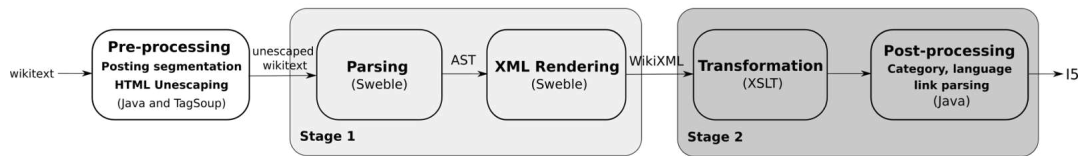


Figure 1: Wikitext to I5 conversion pipeline

Figure 2: Creating a new query to search for an emoji by using the *Query-By-Match* assistant in the annotation view in KorAP

Wikipedia: Linguistic corpus building, exploration and analysis, pages 12–44. Benjamins.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1848–1854. European Language Resources Association (ELRA).

Kupietz, M., Diewald, N., and Margaretha, E. (2020). RKorAPClient: An R package for accessing the German Reference Corpus DeReKo via KorAP. In

Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 7015–7021. European Language Resources Association.

Lüngen, H. and Kupietz, M. (2017). CMC corpora in DEREKO. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and natural Language Processing (CMLC-5+BigNLP)*.

Lüngen, H. and Sperberg-McQueen, M. (2012). A TEI P5 document grammar for the IDS text model. *Journal of the Text Encoding Initiative (jTEI)*, pages 1–18.