Published 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 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 20-24

CMC Corpora in DeReKo

Harald Lüngen
Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
luengen@ids-mannheim.de

Marc Kupietz
Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
kupietz@ids-mannheim.de

Abstract

We introduce three types of corpora of computer-mediated communication that have recently been compiled at the Institute for the German Language or curated from an external project and included in DeReKo, the German Reference Corpus, namely Wikipedia (discussion) corpora, the Usenet news corpus, and the Dortmund Chat Corpus. The data and corpora have been converted to I5, the TEI customization to represent texts in DeReKo, and are researchable via the web-based IDS corpus research interfaces and in the case of Wikipedia and chat also downloadable from the IDS repository and download server, respectively.

1 Introduction

The German Reference corpus DeReKo was started at the Institute for the German Language (IDS) in 1964 and has been continually expanded since then. Currently it contains more than 31 billion tokens and comprises text types as diverse as newspaper text, specialised texts, fiction, speeches and debates, computer-mediated communication and many more.

Though the bulk of DeReKo has always consisted of newspaper/press corpora, we have made new acquisitions in all of the above mentioned genres in the last couple of years (cf. e.g. Kupietz & Lüngen, 2014). In this paper, we would like to introduce three corpora of computer-mediated communication (CMC) that have recently been compiled for DeReKo. CMC is an interesting type of genre that is increasingly used in research on many aspects of language, e.g. interaction, neologisms, or orthography. In the following, we present our Wikipedia corpora in more detail, and in a little less detail the Usenet news corpus, and the Dortmund chat corpus, which make up the three types of CMC corpora currently in DeReKo.

2 Wikipedia

2.1 History

Since the 2000s, Wikipedia corpora have been created in cooperation with the IDS grammar department and have been included in DeReKo. In the first conversion 2005, the German Wikipedia dump, which contains the texts in the WP "wikitext", format was converted to CES, (Corpus Encoding Standard, cf. Ide, 1998), which was used to encode all IDS corpus holdings at that time). Strictly speaking, only Wikipedia talk pages (discussions) constitute CMC, but this first conversion included only the encyclopedia articles.

The 2011 conversion then for the first time included all German talk pages besides the articles and was produced using a new XSLT-based conversion pipeline which converted the wikitext directly into IDS-XCES encoding (Bubenhofer et al., 2011). It was decided that from now on a new Wikipedia conversion for DeReKo should be produced every two years, while the older conversions should always remain a part off DeReKo to enable diachronic analyses and anyway to ensure replicability of analyses. The 2013 conversion was done using an enhanced converter that employed the Sweble parser (Dohrn & Riehle, 2011), which was deemed a more sustainable method for parsing the wikitext format. The parsed wikitext was then passed to XSLT to produce the new target format I5 for DeReKo (Margaretha & Lüngen, 2014). I5 is a continua-

tion of IDS-XCES molded as a TEI P5 customisation which includes new elements, esp. <posting> adopted from Beißwenger et al. (2012), to represent the macrostructure of CMC dialogues as found in the talk pages. The 2013 conversion was characterised 2014 in the DeReKo paper Kupietz & Lüngen (2014). The 2015 conversion is characterised in the following section. Figure 1 gives an overview of the Wikipedia subcorpora in DEREKO sizes over the four conversions.

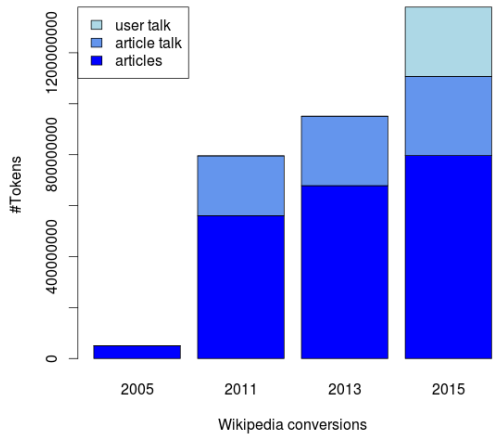

Figure 1: Sizes of WP subcorpora over the years

2.2 Latest WP Corpora

The following is an overview of the new features introduced in our 2015 Wikipedia conversion (which has been a part of DEREKO since 2016):

- User talk pages: Previously, Wikipedia corpora in DEREKO included only the article-related talk pages (where the WP editors discuss the structure and contents of the encyclopedia articles); since 2015 we also included the user talk pages. Every editor in Wikipedia can have a user talk page, and discussions can be conducted on these pages using the wiki software just like with article discussions. Here, users mainly discuss topics not related to the composition of the articles. For example, discussions on article talk pages that become off-topic are sometimes deferred to and continued on a user talk page. The user talk page corpus is of reasonable size similar to the article discussions corpus and constitutes a third Wikipedia corpus within DEREKO (see Fig. 1 and Table 1).
- Language Links: The metadata of the WP article corpus additionally contain all the language links of an article (links to WP pages for the same lemma in different languages). This is useful for creating comparable corpora from different language versions of Wikipedia for cross-lingual analyses. Even the talk pages with discussions about the articles of the same lemma in different languages can be related via these links.
- The I5 <autoSignature> element used in the representation of the discussions now comes with additional type information: "signed", "unsigned", or "user_contribution".
- Timestamps in discussions are now retained in the text (not only in the metadata) and are marked up using the I5 element <timestamp>). When researching WP discussions in COSMAS II, this is the only place where a user can see the date of a user edit.
- The wikitext to I5 converter has been improved, e.g. regarding timestamp identification, posting segmentation, thread identification, and the introduction of properties files for its configuration

	Articles	**Article talk pages**	**User talk pages**
I5 filesize	20G	5.5G	5.2G
#pages	1,802,682	591,460	539,053
#posts	--	6,200,701	5,523,769
#tokens	796,638,747	309,897,027	271,441,322

Table 1: Size of German Wikipedia corpora (conversion 2015) in DeReKo

2.3 Access

All of the above mentioned Wikipedia corpora (conversions from 2005, 2011, 2013, 2015) are included in DEREKO, the German Reference Corpus and can be researched via the COSMAS II, the Corpus Search, Management and Analysis System at the IDS. Presently, no POS annotations are provided for the Wikipedia corpora due to RAM limitations and the way COSMAS II handles annotation indexes. However, with the successor system KorAP (Bański et al., 2013;

	Wikipedia Corpora		
Lang.	**Articles #tokens**	**Article discussions #tokens**	**User discussions #tokens**
de.	796,638,747	309,897,027	271,441,322
en.	2,403,943,177	1,270,217,981	2,698,338,998
fr	764,459,026	131,107,729	372,639,260
hu.	117,987,947	8,293,799	26,215,158
no	99,014,144	5,314,362	32,481,331
es	578,883,431	54,907,258	276,034,367
hr	46,641,724	2,480,966	18,731,167
it	463,022,806	49,826,036	125,573,567
pl	298,207,197	16,558,557	64,126,136
ro	87,117,385	5,195,240	--

Table 2: Overview of sizes Wikipedia corpora in different languages. Interestingly, the German corpora are the only ones where the user discussion corpus is smaller than the articles discussion corpus, and the English corpora are the only ones where the user discussion corpus is bigger than the articles corpus.

Diewald et al., 2016), which has been in public beta test since May 2017, several linguistic annotation layers are already searchable. Besides querying the corpora in COSMAS II and KorAP, corpus, computational linguists can download the I5 files of all Wikipedia corpora from our pub server[1]. The wikitext to I5 converter (java jar files) can also be downloaded from there, complete with documentation (Margaretha, 2015).

2.4 Multilingual Wikipedia corpora

To prove that the conversion pipeline can be used for other Wikipedia language versions, we applied it to convert the Wikipedia dumps for nine further languages which play a role in contrastive and cross-lingual analysis projects at the IDS, see the overview in Table 2. Just like with the German WP, we generated the the corpus types WP articles, article discussions, and user discussions for each language in I5. They, too, are downloadable from our pub server and are even searchable in COSMAS II, where they reside in their own archive separate from DEREKO. However, since COSMAS II is a corpus interface designed for German language corpora, certain COSMAS functions cannot be meaningfully applied to the foreign language corpora, including tokenisation and lemmatization.

2.5 Related Work

The Berlin-based company *linguatools* offers monolingual and bilingual, comparable corpora built from the Wikipedia versions of 23 languages for free download. They are based on Wikipedia dumps from 2014 and contain the complete set of articles available in 2014, but only the articles, i.e. no talk pages. They contain rich metadata, including information about link types of internal and external links, and the WP categories under which an article is subsumed. They are distributed in XML markup and are downloadable from the company website (Linguatools, 2014). The bilingual corpora contain pairs of articles in language A and language B that were linked by Wikipedia language links. There are 23 monolingual and 253 bilingual comparable corpora available.

The linguatools Wikipedia corpus conversions cover more languages and contain somewhat richer metadata than ours. They do not include talk pages, and the XML encoding covers fewer structural phenomena than our I5 encoding. Their bilingual comparable WP-corpora are very useful for cross-lingual or contrastive linguistic analyses. Similar corpora could straightforwardly extracted from our Wikipedia corpora using the language links.

3 Usenet News

While many types of CMC corpora are alternatively identified as *social web corpora*, Usenet newsgroups definitely do not constitute a web genre, let alone a social web one, as the Usenet is based on its own internet protocol called nntp (Horton and Adams, 1987). As a CMC genre, newsgroups work similar to discussion forums, containing user contributions about a common topic organised in threads. Unlike typical Web 2.0 discussion forums, however, the Usenet is non-proprietary i.e. everybody can just use a news client and participate, or even set up their own news server to host newsgroups. Besides, all newsgroups are organised in a single topic hier-

[1] http://www.ids-mannheim.de/direktion/kl/projekte/korpora/verfuegbarkeit.html

archy, i.e. theoretically, for a particular topic, there is exactly one newsgroup. The Usenet started in 1979 and had its heydays in the 1990s, which makes it potentially interesting as a source of more historical, pre-Web 2.0 CMC.

We have compiled a corpus of German with all newsgroups from the news server news.individual.de (run by FU Berlin), with all textual newsgroups from the .de-hierarchy, starting in 2013. The downloaded news messages have been converted to I5 and been annotated with certain microstructural CMC features (Schröck & Lüngen 2015) and are researchable via COSMAS II, but for the time being (as the corpus is not anonymised) only on the premises of the IDS.

Usenet news corpus in DEREKO	
Period	2013-2016
#Newsgroups (all groups in the de. hierarchy)	375
#News messages	1,094,281
#Tokens	92,520,763

Table 3: Usenet news corpus overview

The news corpus in DEREKO is being continued with the latest data from the news server but also to be extended with data from the years before 2013. These, however, would have to be gleaned from a commercial news server.

4 Chat

In a so-called CLARIN-D curation project, the Dortmund Chat Corpus (about one million tokens, cf. Beißwenger et al., 2013) has been prepared for inclusion in CLARIN-D research infrastructures including DEREKO. The project work[2] comprised a conversion to a newly tailored CLARIN-D TEI customisation for chat and other CMC data (Lüngen et al., 2016), CMC-specific part-of speech tagging (Beißwenger et al, 2015), and corpus anonymisation according to the requirements set out in a legal expertise. The result is the Dortmunder Chat Corpus 2.0 as characterised in Table 4.

Dortmund Chat Corpus 2.0	
# log files	470
# posts	131,003
# tokens	1,005,166
File size (CLARIN-D TEI)	100 MB

Table 4: Dortmund Chat Corpus 2.0 Overview

It has been converted to I5 and is integrated in DEREKO and will be searchable through COSMAS II shortly. Besides, it is also available from the CLARIN-D repository at IDS[3].

5 Conclusion

We presented three types of CMC corpora that have recently been compiled at the IDS or curated from an external project and included in DEREKO, the German Reference Corpus, namely Wikipedia (discussion) corpora, the Usenet news corpus, and the Dortmund Chat Corpus. We will continue to build Wikipedia linguistic corpora every two years, i.e. the preparation of the 2017 conversion is impending. It will include a few new features, e.g. new metadata types similar to those available in the linguatools corpora, and also further types of discussion corpora from other Wikipedia namespaces. The Usenet corpora will be updated with the latest data but also be extended with data from the years before 2013. Chat corpora with more recent smart phone chat data will be acquired via a cooperation with the Mobile Communication Database project at the University of Duisburg-Essen.[4] We will also continue to try and compile other types of CMC corpora e.g. from web 2.0 blogs and discussions forums, provided that they come with licenses appropriate for redistribution.

References

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C. and Witt, A. (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Z. and Uszkoreit, H. (eds.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. Poznań: Fundacja Uniwersytetu im. A., 2013. Pages 586-587.

[2] together with partners from the Universities of Mannheim, Duisburg-Essen, and the Berlin-Brandenburg Academy of Sciences.

[3] PID: http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E

[4] http://mocoda.spracheinteraktion.de/

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. Journal of the Text Encoding Initiative, 3.

Beißwenger, M., Ehrhardt, E., Horbach, A., Lüngen, H., Steffen, D., Storrer, A. (2015): Adding Value to CMC Corpora: CLARINification and Part-of-speech Annotation of the Dortmund Chat Corpus. In: Beißwenger, Michael/Zesch, Torsten (Hg.): NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop , September 29, 2015 University of Duisburg-Essen, pages 12-16, German Society for Computational Linguistics & Language Technology (GSCL).

Bubenhofer, N., Haupt, Stefanie, Schwimm, Horst (2011): A Comparable Corpus of the Wikipedia: From Wiki Syntax to POS Tagged XML. Hamburg Working Paper in Multilingualism, 96 B. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-51897

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: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., (eds.): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pages 3586-3591, Paris: European Language Resources Association (ELRA).

Horton, M.; Adams, R. (1987): RFC-1036 Standard for Interchange of USENET Messages. Available online at: http://tools.ietf.org/html/rfc1036 .

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 72–81, New York, NY, USA. ACM.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic

Corpora. In: Proceedings of the First International Language Resources and Evaluation

Conference (LREC), pages 463–470, Granada, Spain.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based standard for linguistic corpora. In: Proceedings of the Second Language Resources and Evaluation Conference (LREC), pages 825–830, Athens, Greece.

Kupietz, M. and Lüngen, H. (2014): Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages. 2378–2385. European Language Resources Association (ELRA).

Linguatools (2014): Wikipedia Comparable Corpora. http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

Lüngen, H., Beißwenger, M., Ehrhardt, E., Herold, A., Storrer, A. (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: Dipper, Stefanie/Neubarth, Friedrich/Zinsmeister, Heike (eds.): Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). Bochumer Linguistische Arbeitsberichte (BLA) 16, pages 156-164, Bochum.

Margaretha, E. and Lüngen, H. (2014): Building linguistic corpora from Wikipedia articles and discussions. In: *Journal for Language Technologie and Computational Linguistics (JLCL)* 2/2014

Margaretha, E. (2015): Documentation of the IDS Wikipedia Converter, 2015. http://corpora.ids-mannheim.de/pub/tools/2015/

Schröck, J. and /Lüngen, H. (2015): Building and Annotating a Corpus of German-Language Newsgroups. In: Beißwenger, Michael/Zesch, Torsten (eds.): NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop , September 29, 2015 University of Duisburg-Essen, pages 17-22, German Society for Computational Linguistics & Language Technology (GSCL).