The Stuttgart DEWAC (SDEWAC)
February 2012

Short description: SDeWaC is a corpus created from a subset of the DeWaC corpus (Baroni/Kilgarriff 2006). It contains 44.084.442 sentences, 846.159.403 word form tokens and 1.094.902 types. The sentences were selected on the grounds of being syntactically parsable with a standard dependency parser for German.

Motivation for the creation of SDeWaC: In the original DeWaC, we found numerous sentences and word sequences which caused problems to our parsers, as they were too long, too short, or simply not syntactically well-formed at all, e.g. (key) word lists. Furthermore, DeWaC countains many duplicates. For such reasons, we decided to isolate the parsable subset of DeWaC. Instead of “corrective cleaning”, problematic data were discarded.

Procedures applied to DeWaC: to prepare SDeWaC, we used pattern matching to identify material not usable for parsing (cf. Quasthoff et al. 2006). The remaining sentences were parsed with the FSPar dependency parser by Schiehlen (2003).

This parser assigns numbers to nodes in the dependency structure, and whenever a node cannot be localized within a dependency structure, a “0” is assigned. By calculating an error rate (the number of nodes with “0” assigned, related to the length of the sentence) and by selecting sentences with low error rates only, we assume to only find sentences that FSPar can parse. Note that this method has not yet been evaluated; a paper describing the method in detail and an evaluation are planned for mid 2012. Faaß et al. (2010) describes parts of the cleaning process.

The corpus is distributed as version 03 in two formats. A separate document (file “web-adress-list.txt”) contains the details of the URLs of the source texts; in the corpus, sources are numbered; “-” is assigned to unknown sources; numbers have up to 5 digits.

Note that version 03 has been parsed at the University of Stuttgart with Bernd Bohnet's dependency parser (Bohnet (2010), contact Kerstin Eckart for more information), however, as it still contains a number of unusable sentences, further cleaning is ongoing (expect version 04 to follow).

Formats:

Format 1: One sentence per line. In front of the sentence, there are three columns (separated by tabs):

<year=”...”/> <source=”....”/> <error=”....”/>

For example:

<year="2007"/> <source="10475"/> <error="0"/> Und wie funktionieren eigentlich Atomuhren ?

Format 2: One token per line, tokenized with Schmid's tokenizer (Schmid (1994a)) and POS-annotated/lemmatized with Schmid's tree-tagger (Schmid (1994b)), like the original DeWAC. Here, the metadata are added in a pseudo-xml structure <sentence>, as shown in the following example:

```
<sentence>
<year>="2007"/>
<source="10475"/>
<error="0"/>
<s>
Und KON und
wie KOUS wie
funktionieren VVFIN funktionieren
eigentlich ADV eigentlich
Atomuhren NN Atomuhr
? $. ?
</s>
</sentence>
```

References

Baroni/Kilgarrif 2006. Marco Baroni and Adam Kilgarriff. Large linguistically-processed Web corpora for multiple languages. Conference Companion of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics).2006. 87 – 90.

Bohnet (2010). Bernd Bohnet. Top accuracy and fast dependency parsing is not a contradiction. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 89–97, Beijing, China, August 2010.

Faaß et al. (2010). Gertrud Faaß, Ulrich Heid, and Helmut Schmid. Design and application of a Gold Standard for morphological analysis: SMOR in validation. In Proceedings of the seventh LREC conference , pages 803 – 810, Valetta, Malta, May 19 – 21 2010. European Language Resources Association (ELRA).

Quasthoff et al. (2006). Uwe Quasthoff, Matthias Richter, and Christian Biemann. Corpus portal for search in monolingual corpora. In Proceedings of the LREC 2006, Genova, Italy, 2006.

Schiehlen (2003). Michael Schiehlen. A cascaded finite-state parser for German. In Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03), Budapest, April 2003.

Schmid (2000). Helmut Schmid. Unsupervised Learning of Period Disambiguation for Tokenisation, 2000. Internal Report, IMS, University of Stuttgart.

Schmid (1994b). Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In International Conference on New Methods in Language Processing, pages 44 – 49, Manchester, UK, 1994.