The Fragment Corpus (FraC)

Eva Horch and Ingo Reich
(Universität des Saarlandes, Germany)

We present the Fragment Corpus (FraC), a corpus for the investigation of fragments (see Morgan 1973), i.e. incomplete sentences, in German. The corpus is a mixed-register corpus and consists of 17 different text types including written (newspaper texts, legal texts, blogs etc.) and spoken texts (dialogues, interviews, radio moderations etc.), as well as social media texts (tweets, sms, chats). Each subcorpus comprises approx. 2000 utterance units (including fragments, following the orthographic notion of sentence) which amounts to a total corpus size of 380K tokens. The data was taken from electronically available sources.

Depending on availability, metadata include information on source, author, date and text type. The corpus is annotated for tokens, lemmas and POS using TreeTagger (Schmid 1994, 1995). In addition, each utterance unit is manually annotated with:

- its syntactic category (like VP, NP or SEQUENCE for sequences of syntactic phrases)
- a listing of immediate constituents (in the case of sequences or coordinations)
- the discourse function of the utterance unit (like greeting, listing, discourse topic)
- idiosyncratic information (like 'containing an acronym', 'formulaic expression')
- omission type (like topic drop, object drop, article and copula omission)
- null elements (for not realized elements in case of article and copula omission)

FraC was built within the project B3 of the CRC 1102 on *Information Density and Linguistic Encoding* (IDeaL). In this project, the corpus is primarily used for corpus linguistic studies of fragments as well as for building language models in order to investigate the actual use of fragments by relating their use to information-theoretic notions like ID (see Shannon 1948). It is planned to also annotate the corpus with Surprisal values for each token based on trigrams and to parse the corpus (with the Stanford parser, Rafferty & Manning 2008).

The source of the corpus is in XML format. The corpus is encoded in two different formats: (i) in CQP format (cf. IMS Open Corpus Workbench CWB; Evert & Hardie, 2011) for query and statistical analysis and (ii) in ANNIS format (Krause & Zeldes 2016). The corpus will be made accessible through the ANNIS multilayer corpus tool and the CQPweb installation of Saarland University under the current copyright laws. In the medium term, we would like to make FraC available through the CLARIN-D repository of Saarland University.

References

Krause, T. & A. Zeldes (2016). ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities 2016* (31).

Morgan, J. L. (1973). *Sentence fragments and the notion 'sentence'*. In B. Kachru, R. Lees, Y. Malkiel, A. Pietrangeli, & S. Saporta (Eds.), Issues in Linguistics: Papers in Honor of Henry and Renée Kahane (pp. 719-751). University of Illinois Press.

Rafferty, A. & C.D. Manning (2008). *Parsing three German treebanks: lexicalized and unlexicalized baselines*. In ACL Workshop on Parsing German.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International*

Conference on New Methods in Language Processing, 44–49. Manchester, UK.
Schmid, H. (1995). Improvements in part-of- speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.
Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27(4), 623-656.