Testing Statistical Measures for Morphological Analysis

Petra C. Steiner

Lexical productivity, especially by compounding and derivation, is a characteristic of German word formation. This leads to bottleneck problems in different fields such as the building of terminology or Information Retrieval. Most tools for the analysis of German word forms are restricted to flat morphological structures (e.g. SMOR by Schmid (2004), Gertwol by Haapalainen & Majorin (1995), MORPH by Hanrieder (1991, 1996), and TAGH by Geyken & Hanneforth (2006). The same holds for experimental settings (e.g. Cap 2014, Koehn & Knight 2003). Only Würzner & Hanneforth (2013) tackle the problem of full morphological parsing, restricted to adjectives, by using a probabilistic context free grammar for parsing.

For detailed semantic processing, the recognition of hierarchical word structures is prerequisite. Here a challenge arises in that word forms are amenable to ambiguous structure interpretations. Probabilistic context free grammars build on frequencies of constituents but they do not exploit information on co-occurrences and positions of morphological units. By contrast, the methodological framework of this investigation builds on the hypothesis that morphological analysis of word tokens can be improved by the specific contextual information.

Steiner & Ruppenhofer (2015) and Steiner (2016) developed a method for building parts of morphological structures by using the results from SMOR (Schmid 2004) and reducing the set of all possible low-level combinations by weighting measures. The contextual information was broadly defined as the lexical inventory of the Mannheim corpus (see Gulikers et al. 1995, 102ff.).

The current investigation tests different weighting measures and definitions of contextual information. The leading hypothesis is that for the different levels of morphological analyses different weighting measures are suitable.

The frequencies for the measures were drawn from two sources: (a) the tokenized items of the Korpus Magazin Lufthansa Bordbuch (MLD), which is a part of DeReKo-2016-I (IDS 2016, Kupietz et al. 2010) for the contextual information (b) the CELEX database for German (Baayen et al. 1995) for the frequencies of those constituents whose frequencies could not be drawn from the corpus, including also derivational affixes. The frequency values from CELEX are adjusted according to the size of corpus and texts. As in previous investigations, weighting with geometric means score lead to typical errors in the analyses, whereas measures based on frequency and length of linguistic units yield better results.

References

Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. *The CELEX lexical database (CDROM)*. Linguistic Data Consortium, Philadelphia, PA.

Cap, Fabienne. 2014. *Morphological processing of compounds for statistical machine translation*. Dissertation. Universität Stuttgart. [http://elib.uni-stuttgart.de/opus/volltexte/ 2014/9768].

Geyken, Alexander & Thomas Hanneforth. 2006. TAGH: A Complete Morphology for German based on Weighted Finite State Automata. Yli-Jyrä, Anssi, Lauri Karttunen & Juhani Karhumäki, ed. *Finite State Methods and Natural Language Processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*. Berlin/Heidelberg, Springer. 55-66.
[http://www.dwds.de/static/website/ publications/text/Geyken_Hanneforth_fsmnlp.pdf].

Gulikers, Léon, Gilbert Rattink & Richard Piepenbrock. 1995. German Linguistic Guide. Harald Baayen, Richard Piepenbrock, and Léon Gulikers, ed. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, Philadelphia, PA.

Haapalainen, Mariikka & Majorin, Ari. 1995. GERTWOL und morphologische Disambiguierung für das Deutsche. *Proceedings of the 10th Nordic Conference on Computational Linguistics, Helsinki, Finland.*

Hanrieder, Gerhard. 1991. *Robustes Wortparsing. Lexikonbasierte morphologische Analyse*

(komplexer) deutscher Wortformen. Universität Trier. Master Thesis.

Hanrieder, Gerhard. 1996. MORPH - Ein modulares und robustes Morphologieprogramm für das Deutsche in Common Lisp. Hauser, Roland, ed. *Linguistische Verifikation. Dokumentation zur Ersten Morpholymics 1994*. Tübingen: Niemeyer. 53-66.

Institut für Deutsche Sprache. 2016. *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2016-I (Release from 31.03.2016)*. Mannheim: Institut für Deutsche Sprache. [www.ids-mannheim.de/DeReKo.]

Koehn, Philipp & Kevin Knight. 2003. Empirical methods for compound splitting. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*. 187-193.

Kupietz, Marc, Cyril Belica, Holger Keibel, Andreas Witt. 2010. The German Reference Corpus DeReKo: A primordial sample for linguistic research. Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, Mike, Daniel Tapias, ed. 2010. *Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010)*. S. 1848-1854.

Schmid, Helmut, Arne Fitschen & Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal. 1263-1266. [http://www.cis.uni-muenchen.de/~schmid/papers/smor.pdf].

Steiner, Petra (2016). *Kontextbasiertes morphologisches Parsing*. Poster presented at the DGfS-CL Postersession of the Annual Meeting of the DGfS, University of Konstanz, February 24-26.

Steiner, Petra & Josef Ruppenhofer (2015). Growing Trees from Morphs: Towards Data-Driven Morphological Parsing. Bernhard Fisseni, Bernhard Schröder & Torsten Zesch, ed: *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30th September - 2nd October 2015*. 49–57. [http://gscl2015.inf.uni-due.de/wp-content/uploads/2016/02/GSCL-201508.pdf]

Würzner, Kay-Michael & Thomas Hanneforth. 2013. Parsing Morphologically Complex Words. *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, FSMNLP 2013, St. Andrews, Scotland, UK, July 15-17, 2013*. 39-43. [http://aclweb.org/anthology/W/W13/W13-1807.pdf]