Influence of information density on deletion rates in German

Erika Brandt, Frank Zimmerer, Bistra Andreeva, Bernd Möbius

Universität des Saarlandes, Sprachwissenschaft und Sprachtechnologie

ebrandt@coli.uni-saarland.de

This study investigated the influence of information density on the deletion of segments in German content words. As controls, a prosodic models containing primary lexical stress, articulation rate, and boundary, as well as speech sound class and word frequency were used in the statistical analysis. We hypothesized that deletion rates increased with higher predictability while controlling for known effects of word frequency and prosodic factors on deletion rates (Cohen Priva, 2015), and that this effect held for predictability estimated on the preceding and the following context of the phoneme.

The Siemens Synthesis corpus (SI1000P) was analyzed which contains newspaper read speech from two professional male speakers (Schiel, 1997). Segment deletions were defined based on the canonical transcriptions contained in the corpus materials: If there was a discrepancy between the expected phoneme string and the uttered phoneme string, this was marked by trained annotators.

Information density was estimated as surprisal ($S(unit_i) = -\log_2 P(unit_i|context)$). Surprisal values for the preceding and following context were calculated from a bi-phone language model using SRILM (Stolcke, 2002). As a text corpus for language modeling and word frequency counting sDeWaC was preprocessed and normalized using German-Festival (Samlowski, 2016). Articulation rate was defined as phonemes per second excluding pauses. Speech rate was binned into slow, mid and fast speech rate. Primary lexical stress with factor levels unstressed and stressed was based on the canonical transcription of the Siemens Synthesis corpus. Boundary was defined as word, phrase or no boundary. Statistical analyses with GLMMs were performed for the entire data set and separately for /ə/ deletion and /t/ deletion. For the /ə/ model we included deletion predicted by the *Duden Aussprachewörterbuch* (Dudenredaktion, 2015) as an additional predictor, while following context with neutralizing, other consonantal, and vocalic context was taken into account in the /t/ model.

Easily predictable segments were more likely to be deleted than segments that were difficult to predict. This was true for both preceding and following context. High frequent words had higher deletion rates than low frequent words. At mid and slow speech rate, segment deletion rate was significantly lower than at fast speech rate. Segments in unstressed syllables were more likely to be deleted than segments in stressed syllables. At phrase and word boundary there were significantly less deletions than at no boundary position. In the /t/ model, following context was a strong predictor of deletion, while surprisal of the following context and stress were not. Low surprisal of the preceding context, however, predicted /t/ deletion. Duden rules and low surprisal of the following context predicted /ə/ deletion, whereas surprisal of the preceding context and speech rate were insignificant in the model.

References

Cohen Priva, U. (2015). Informativity affects duration and deletion rates. *Laboratory Phonology, 6*.

Dudenredaktion (2015). Duden – Das Aussprachewörterbuch. 7. Volume.

Samlowski, B. (2016). The syllable as a processing unit in speech production: evidence from frequency effects on coarticulation. Phd thesis, Bielefeld: University of Bielefeld.

[dataset] Schiel, F. (1997). *Siemens Synthesis Corpus - SI1000P*. University of Munich. https://www.phonetik.uni-muenchen.de/Bas/BasSI1000Peng.html.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit, In: *Proceedings of the International Conference of Spoken Language Processing*, Vol. 2., Denver, 901--904.