Broad Phonetic Transcription

Steven Gillis

The complete CGN corpus will be enriched with a "broad phonetic" or "phonemic" transcription. The term "phonemic" covers the exact scope of the linguistic annotation at this level: the aim is to arrive at a transcription that relates the sounds produced by the speakers to the phonemes of Dutch. Hence, marking of common processes such assimilation are considered only in so far as they concern phonemic distinctions (e.g., nasal assimilation is transcribed only if the targets are the phonemes /m, n, N/). Moreover no diacritic marking of idiosyncrasies in a speaker's pronunciation (such as diphthongization, palatalization and the like) is marked. Only phenomena that clearly cross phoneme boundaries are considered. The transcription is meant to be phonemic and not morphophonemic, which implies that phonological processes such as final devoicing will be considered in the transcription (e.g., <hoed> 'hat' will normally be transcribed as /hut/ and not as /hud/, the latter being morphophonologically motivated because of the plural form /hud@n/).

The symbol set used in this task is very close to the one developed by SAMPA, and akin to DISC – utilized in CELEX. Divergences from these two phonetic conventions are well motivated in a Gillis (2002). Primarily they relate to avoiding the use diacritics such as length marks and the use of a transparent set of symbols that is more phonologically motivated and less phonetically grounded.

The transcription adheres to the word as a central unit. Hence, the main guiding principle states that a one-to-one relationship is established between the central orthographic tier and the phonemic one, as is also the case for other linguistic annotations such as part-of-speech tagging and syntactic annotation. One of the main problems induced by this way of conceptualizing the relationship between the orthography and the phonemic transcription is the problem of cross-word assimilations, such as the degemination in <om meer> 'for more' (e.g. in <hij vraagt om meer> 'he asks for more'). In orer to deal with thesefairly frequent cases, an intricate though transparent notation was developed for marking that a sound is shared by two or more consecutive words. These conventions are led down in a fully explicit protocol (Gillis et al. 2001).

The 'broad phonetic transcription' is arrived at in the following way: on the basis of the orthographic tier a phonemic representation is computed in a fully automatic way, using the TREE-TALK program developed by Daelemans & Van den Bosch (2001) and implemented in TiMBL (Daelemans et al. 2001). This machine learning system is trained on the phonemic transcriptions represented in CELEX(for the Northern Dutch variant) and on FONLIEX (which contains three variants of Southern Dutch). Subsequently the transcriptions were checked and if necessary corrected by research assistants who were properly trained on accomplishing this task, using the CGN protocol and a dedicated set of training materials. The procedure leading up to the phonemic transcription was supervised by a phonetically trained research assistant who performed a final check on all outputs.

This procedure is followed for 10% of the CGN data. The rest of the data will be automatically transcribed using the procedure outlined above, with a retraining of the automatic transciber including the manually verified materials.

References

Daelemans, Walter and Antal van den Bosch., TreeTalk: Memory-Based Word Phonemisation.. R. I. Damper (Ed.) Data-Driven Techniques in Speech Synthesis. Kluwer Academic Publishers, 149-172, 2001

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide. ILK Technical Report 01-04, 2001

Gillis et al. (2001) Protocol

Gillis (2002) Motivering protocol fonemische transcriptie