BAS Infrastrukturen zur Technischen Sprachverarbeitung (BITS)

Teilprojekt 8 (Doku 8/6e)

Construction of the Unit Selection Corpus

Content: The construction of the Unit Selection Corpus is described. Mainly it is an explanation from Norbert Braunschweiler from the IMS in Stuttgart who provided the corpus for us.

Author: Tania Ellbogen
Date: 05.04.2006
Version: 1.1

Construction of the Unit Selection Corpus

The corpus was constructed by Norbert Braunschweiler from the IMS in Stuttgart who provided it to the BITS project. The main point collecting the sentences was a complete set of German diphones plus three French phones.

Mail from Norbert Braunschweiler:

The sentences were mainly chosen by a program whose aim was to find every possible diphone combination for German (plus the three French vowel on, en, an). The publicly available program was written by Alberto Sesma from the university of Barcelona. In short the program calculates the number of required sentences (or words) to achieve the desired coverage starting from a given corpus that is phonetically transcribed. The input corpus was a subcorpus from the TAZ-corpus with about 170000 sentences, annotated with POS-tags from Stuttgart. These sentences were automatically corrected. Then the sentences were automatically transcribed.

Partly the transcription was automatically corrected by means of scripts.

The result was a file with known coverage of diphones which was completed manually with the still missing diphone combinations. This was achieved by adding invented SUS-sentences (semantically unpredictable sentences [...]). Thus we could achiecve a complete coverage of diphones for German plus combinations of the three French vowels.