The Database

The Kiel Corpus of Spontaneous Speech

Benno Peters

1 Introduction

In two German Research Council (DFG) funded projects on *Sound Patterns* and *Prosodic Phrasing in German Spontaneous Speech* (Ko 331/22-1,2 and Ko 331/23-1), the setting up of the *Kiel Corpus* database has been completed. Its now the largest orthographically, segmentally, and prosodically annotated database of a single language worldwide, including read speech (IPDS 1994), besides the data taken from (quasi-)spontaneous speech dialogues (IPDS 1995, 1996, 1997), which were analysed within the above-mentioned projects.

2 Datcollection

For the collection of the spontaneous speech data, two scenarios were used:

- In the *appointement-scheduling* scenario two dialogue partners make various appointements on the basis of calender sheets and academic timetables. Most of the speech data elicited by means of this scenario were recorded within the VERBMOBIL project (Karger and Wahlster 1994) with a technique that prevents the recording of dialogue partners speaking simultaneously. When a speaker keeps a button pressed, his/her own speech signal is recorded, at the same time blocking the other speaker's channel. This method was chosen in the VERBMOBIL project for technical reasons of easy data processing, although it heavily affects natural dialogue control between speakers.

- The *Video Task* scenario, also referred to as *Daily Soap Scenario*, was developed at the IPDS, specifically for speech data collection in dialogue (Peters 2001). In this scenario, similar but non-identical video material is presented to two subjects sitting in separate rooms. After the presentation, the subjects discuss differences and similarities of what they have seen and heard. For the data collection carried out so far two tapes were spliced together from a number of episodes of the well-known German television series LINDENSTRASSE. The video sequences on each tape cover approximately 15 minutes and diverge partly as to selection, sequence and completeness of single scenes. The method of *Video Task* data collection allows parallel speech recording of both dialogue partners. Thus, natural interaction between speakers is not constrained. These data are the basis for investigations of dialogue mechanisms in German spontaneous speech, reported in contributions to this volume.

3 Quantity of collected speech data

16 dialogues in total (each of them comprising seven subdialogues) from 6 female, 8 male, and 2 mixed pairs in the *appointement-scheduling* scenario were edited and published as *The Kiel Corpus of Spontaneous Speech Vols. I-III* (but without prosodic labels). The 16 dialogues have a total duration of approximately 240 minutes (approx. 37,000 consecutive words).

In addition, six dialogues were recorded from 4 female and two male pairs by means of the *Video Task* scenario. The duration amounts to 80 minutes (approx. 13,000 consecutive words). These dialogues will be published shortly.

Volumes I and II of *The Kiel Corpus of Spontaneous Speech* as well as the six dialogues of the *Video Task* scenario served as the database for the contributions presented in this volume.

4 File names and data processing

The *appointment-scheduling* dialogues are referenced by *g<3digits>a<3-digits>*, the first letter *g* referring to the dialogue type (*appointment scheduling*), the second letter *a* to the place of data collection (Kiel), the first set of digits to the dialogue and its subsessions, the second set of digits to the turns inside a dialogue subsession. The LINDENSTRASSE dialogues have *l* for *g* and three-letter speaker IDs for the second digits in the single-speaker files.

The processing of the speech data comprises the following steps:

- Orthographic ***transliteration*** including several special characters representing phonetic phenomena such as breathing, pauses and the like.

- Automatic generation of a phonematic ***transcription*** using the graphem-to-phonem-module of the text-to-speech-system RULSYS (Kohler 1997).

- ***Segmental labelling*** on the basis of the automatically generated phonematic transcription.

- ***Prosodic labelling*** using the symbolic system PROLAB, which is based on the *Kiel Intonation Model* (*KIM*) (Kohler 1991).
- Creation of commentary files for suprasegmental phenomena that are not captured by the prosodic labelling. They include changes in voice quality, F0 range, and intensity. The commentary files have only been created for the *Video Task* data.
- Labelling of dialogue structure in relation to different types of turn change-overs using a newly developed labelling system, which differentiates turn-internal and turn-final prosodic boundaries, and among the latter overlapping and non-overlapping turn transitions (only for *Video Task* data).
- Automatic transformation of the transliteration and labelling files into the KIELDAT data bank format, which provides quick access to phonetic structures at the label and acoustic levels with the help of data bank functions.

A large part of the work on the corpus was carried out by student assistants under the guidance of research staff. The transliteration and the segmental labelling conform to the conventions set out in Kohler, Pätzold, and Simpson (1995) . The inital training in prosodic labelling took place by means of interactive training materials accompanying KIM and PROLAB (Peters and Kohler 2004). The research included the completion of the prosodic labelling of the data in both scenarios and of the commentary files for the data in the *Video Task* scenario. So, the entire collected corpus is now available with orthographic, segmental as well as prosodic annotations – *read speech*, *appointment-scheduling* VERBMOBIL and *Video Task* (LINDENSTRASSE), and will be published in a new edition on CD-ROM shortly.

Literatur

IPDS (1994). *The Kiel Corpus of Read Speech*, Volume 1, CD-ROM#1. Kiel: IPDS.

IPDS (1995). *The Kiel Corpus of Spontaneous Speech*, Volume 1, CD-ROM#2. Kiel: IPDS.

IPDS (1996). *The Kiel Corpus of Spontaneous Speech*, Volume 2, CD-ROM#3. Kiel: IPDS.

IPDS (1997). *The Kiel Corpus of Spontaneous Speech*, Volume 3, CD-ROM#4. Kiel: IPDS.

Karger, R. and W. Wahlster (1994). *VERBMOBIL Handbuch*. Verbmobil Technisches Dokument Nr. 17. Saarbrücken: DFKI.

Kohler, K. J. (1991). A model of German Intonation. In K. J. Kohler (Ed.), *Studies in German Intonation*, AIPUK 25, pp. 295–360. Kiel: IPDS.

Kohler, K. J. (1997). Modelling prosody in spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pp. 187–210. New York: Springer.

Kohler, K. J., M. Pätzold, and A. Simpson (1995). *From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech*. AIPUK 29. Kiel: IPDS.

Peters, B. (2001). *'Video Task' oder 'Daily Soap Szenario': Ein neues Verfahren zur kontrollierten Elizitation von Spontansprache*. URL: http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/Linda21.html.

Peters, B. and K. J. Kohler (2004). *Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM*. URL www.ipds.uni-kiel.de/kjk/forschung/lautmuster.en.html.