LABELLED DATA BANK OF SPOKEN STANDARD GERMAN THE KIEL CORPUS OF READ/SPONTANEOUS SPEECH

Klaus J. Kohler

Institut für Phonetik und digitale Sprachverarbeitung
Universität Kiel, Germany

ABSTRACT

This paper outlines the successive steps in the setting up of a labelled data bank of German read and spontaneous speech at IPDS Kiel.

1. INTRODUCTION

Large computer accessible speech data bases are a prerequisite in modern studies of the phonetic (segmental and prosodic) regularities of individual languages, their dialectal varieties and their different speaking style realizations. This holds for basic phonetic and phonological research as well as for its speech technology application, e.g. in the training of speech recognition systems. However, the speech wave recordings themselves are not sufficient for any scientific investigation; for the data to be retrievable they have to be annotated. In its simplest form this may be an orthographic representation of the spoken words.

In the Verbmobil Project [6.] two constraints were imposed on a German data collection:

- spontaneous speech in an appointment scheduling scenario
- orthographic transliteration, including the representation of non-lexical material.

Phonetic segmentation and labelling as well as prosodic markings, in addition to orthographic symbolization, were accorded a much lower importance, i.e. they were only to be provided for a small subsection of the total recorded data base. What was, however, stipulated was the automatic generation of canonical phonemic-type transcriptions for all the words in the corpus, compiled in a canonical pronunciation lexicon. As work proceeded the speech recognition groups within the project, even the most ardent adherents of a purely statistical approach, requested the listing of the most frequent actually occurring pronunciation variants in a variants lexicon to assist in the training of their systems.

It was on the strength of these demands from within the Verbmobil group and of our basic conviction that progress in automatic speech recognition will be enhanced by phonetically labelled data that we issued a more extensive labelled 'Kiel Corpus of Spontaneous Speech' on CD-ROMs (2 so far [4.,5.]). These CD-ROMs only include dialogues that have been manually segmented and labelled in respect of their actual pronunciation, i.e. beyond their canonical lexical citation form representation. The Kiel Corpus thus differs in basic orientation from the Bavarian Archive for Speech Signals (BAS), cf. [21.], and it exactly parallels the identically structured previous 'Kiel Corpus of Read Speech' [3.]. The two data bases furthermore allow a comparison of speaking styles, as a contribution to the generally relevant question "how are German words pronounced?".

2. FROM RECORDING TO TRANSLITERATION

2.1. German Read Speech

The 'Kiel Corpus of Read Speech' on CD-ROM#1 comprises the following data. In total 598 sentences and 2 stories, containing 4932 word tokens and 1671 word types, were recorded in a sound-treated room, using a condenser microphone Neumann U87. Subjects read the sentences from monitor prompts, one at a time, the texts from a board. For each read sentence and each read text (section), ASCII-coded and stored in individual text files, a separate digital signal file (16 bit/16kHz) was created. Of 53 speakers (27 male, 26 female) 2 read the entire corpus, 3 a subset of 200 sentences, and 48 either subsets of 100 sentences or one of the two stories. The number of recorded words totals 31,374. For further details see [7.,22.,23.].

2.2. German Spontaneous Speech

Following the original Verbmobil guidelines recordings have been carried out in quiet rooms with two dialogue partners communicating via headsets (Sennheiser HDM 410 or 414). The recording environment ensures high quality speech and good channel separation. Speakers have to press a button whenever they wish to speak to their partner. Only when the button is pressed, which is signalled by a green lamp lighting up, can a speaker be heard by the dialogue partner and recorded. The pressing of the button also blocks the other speaker's channel. This set-up leads to a strict delimitation of turns.

The speech signals are recorded directly to hard disk into a multiplex stereo file (2x16bit/16kHz), on a PC AT486/66 platform with about 500MB disk space, sufficient for recording sessions in excess of one hour. A backup on DAT is produced at the same time. A DSP (Loughborough LSI96002 board) controls the I/O channels. For each dialogue there is a single file which is subsequently demultiplexed and split automatically into two files, one for each channel. The automatic splitting of each of these into separate turn files is made possible by recording constant known signal markers onto the other channel while a speaker is holding the button pressed [18.].

The data acquisition platform is kept so flexible as to allow the realization of other scenario constraints in the data recording and processing, such as overlapping dialogues (without button pressing). In order to get as much material as possible from a single speaker each dialogue session is divided into subsessions. The data included in the Kiel Corpus on the CD-ROMs published so far were obtained in a scenario of eight subsessions, the first of which is used for familiarization with equipment and task and for setting the recording

level, but is excluded from the Corpus. In each subsession the subjects are given fresh sets of calendar sheets and an academic time table, with different shaded areas (representing unavailability) for each of the two dialogue partners. The task in each case is to make appointments of a prespecified nature. For further details see [16.].

The 'Kiel Corpus of Spontaneous Speech' on CD-ROM [4.,5.] includes 82 dialogue subsessions from 16 speaker pairs, 19 male and 13 female speakers. The numbers of word types and tokens in the entire Corpus so far total 1,587 and 25,603, respectively.

Whereas in the case of read speech the recording succeeds an orthographic model, the procedure is reversed in the acquisition of spontaneous speech: the speech recordings have to be transliterated orthographically post hoc. (To reserve the term transcription for phonetic symbolization, an orthographic rendering of speech through spelling conventions is referred to as transliteration.) With regard to lexical material and punctuation the spelling conventions of DUDEN [24.] are applied, but they have had to be extended to cover non-lexical and paralinguistic phenomena, such as hesitations, pauses, breathing, dysfluencies, external noises etc. The extended alphabet in ASCII coding as well as the rules for its use are set out in [16.]. The turns of one dialogue subsession are transliterated in a single text file, using an interactive computer environment.

3. FROM CANONICAL TRANSCRIPTION TO PHONETIC VARIANCE

The further speech processing stages are identical, irrespective of read or spontaneous data.

3.1. Generating Canonical Transcriptions

In all cases the orthographic text files are automatically converted into segmental phonemic transcription using the grapheme-to-phoneme module and the pronunciation exceptions lexicon of the RULSYS/INFOVOX German TTS system [1.,14.]. To cope with the expanded symbolic repertoire of spontaneous speech transliterations the transformation rules, originally devised for standard orthographic text, have had to be supplemented. The alphabet used is modified and augmented SAMPA [16.]. For each orthographic text file input the module automatically generates a transcription file output. It is manually corrected (appr. 3% error rate for running text) and represents a lexical citation form pronunciation: *canonical transcription.*

Another program combines corresponding orthographic and canonical files and, in the case of spontaneous data, brings them in line with the speech files, i.e. splits them up into turn size. The resulting prototype label file contains the file name (corpus and speaker references), the orthographic text and the canonical transcription corresponding to a speech signal file, as well as the list of phonemic labels, one per line, taken from the canonical transcription. This prototype label file is the basis for subsequent manual labelling (see **Table 1**).

3.2. Phonetic Segmentation and Labelling

A prototype label file and the corresponding speech signal file are input to an adaptation of the KTH/Stockholm MIX program [16.], linking the two and generating a label file of the actual pronunciation (see the example in **Table 1**):

- Through visual inspection of speech wave and spectrogram displays, supported by auditory control, the phonemic labels, taken one after another from the prototype label file list, are manually aligned with the initial time marks of corresponding speech signal segments.
- The speech segment with this label extends as far as the time mark at the beginning of the next segment.
- The segmentation is thus strictly linear without overlap.
- All the canonical labels are kept, and may be augmented and modified. There are 4 possibilities: acceptance (S), replacement (S-S'), deletion (S-), insertion (-S').
- In the case of deletion the next label is aligned to the same time mark as the label for the deleted segment; it thus receives zero duration.

For further details on segmentation and labelling see [16.].

```
g071a008.s1h
TIS008.
ja , aber immer . dann<Z> haben wir das +/auf je ma=/+ jedenfalls mal geklärt .
oend
j 'a:  ,  Q a: b 6+  Q 'I m 6  .  d a n+ z:  h a: b @ n+  v i: 6+  d a s+  Q aU f+  j e:+  m a: =/+  j 'e: d @ n #f "a l s  m a: l+  g @ k l 'E: 6 t.
kend
c:  -h:  j 'a:  ,  Q -q a: b 6+  Q- -q 'I m 6  .  c:  -p:  %d -h a-@ n+ z:  -:k -p:  h a: b-m @- n-+  v i:6+  d -h a s+  Q- -q aU f+  j e:+  m a: -l =/+  -p:  j 'e: d-n @- n- #f "a l s  m a: l+  g -h @- k -h l 'E:6 t -h  .
hend
1249 #c:
1249 #-h:
4050 ##j
5183 $'a:
7492 #,
7492 ##Q
8236 $-q
8236 $a:
9225 $b
9955 $6+
11181 ##Q-
11181 $-q
11181 $'I          . . .
```

Table 1: Example of a label file.

4. FROM SEGMENT TO PROSODY

The linear segmental phonemic frame allows the systematic and economical representation of lexical items in canonical citation forms and the labelling of actual pronunciations with reference to them. This makes it possible to search large labelled data banks very efficiently for phonemic-type modifications, such as assimilations and elisions, in connected speech. But this linear segmental phonemic approach excludes important suprasegmental aspects of two types

- articulatory features that can no longer be linked to single phonemes
- utterance prosody: prosodic phrasing, stress, intonation, speech rate, register.

4.1. Non-linear Components of Articulation

Canonical segments may not be discernible as such in the actual

speech signal and will therefore have to be marked as deleted in labelling. But reflexes may still be present as componential modifications of the remaining segment strings, referable to such processes as glottalization, nasalization, palatalization, velarization etc. For example, in *´könnten´* (canonically **k9nt@n+**, see [12.,16.]) the first nasal consonant may be deleted as a sequential element, but a residue of nasalization still linked to the preceding vowel as a componential feature. Furthermore, the plosive **t** may be realized as glottalization somewhere in the sonorant context, without a precise temporal and segmental alignment. In both cases the articulatory components require a non-linear symbolization, i.e. markers that do not receive durations: **=-~** and **t-q** in labelled **kh''9=-~n-:t-q@-:n+** refer to nasalization and glottalization and are aligned at the same point on the time scale as the following non-deleted segment **n**, indexing phonetic parameters in the environment (for details see [16.]). [SOUND A988S01.WAV] [IMAGE A988G01.GIF]

In other cases the componential reflexes of deleted segments are more complex to specify phonetically and are symbolized by a cover label **MA**, i.e. a general "marker" preceding deleted symbols to indicate some phonetic residue (for details see [15.,16.]). **MA** again refers to some contrastive feature that distinguishes the pronunciation actually found from the one represented simply by deletions. In the signal the articulatory component is located to the left and/or the right of the marker. [SOUND A988S02-4.WAV] [IMAGE A988G02-4.GIF]

If in these instances only the segmental deletions were marked there would be a loss of contrastive phonological information because the signal contains more relevant phonetic features. But the strictly linear segmental phonemic approach is not able to represent this distinctivity. So it needs supplementing by non-linear elements for an adequate phonological account of speech, and the labelling of data bases has to take this requirement into consideration. The Kiel Corpus is built on these principles, combining, in a complementary phonology [9.], the advantages of linear segmental canonical base forms for lexical data bank searches with the need for contrastive phonetic adequacy.

4.2. Utterance Prosody

Prosodic features in the traditional sense of the term, i.e. at the utterance level, need to be included in speech corpus labelling for two reasons:

- as contextual frames for investigations at the segmental and componential articulatory levels
- as fields of study in their own right.

In the Kiel Corpus, prosodic labels, indexed by the special marker **&**, are aligned to the speech wave on the same tier as all the other labels to make cross-references as easy and flexible as possible. The allocation of prosodic labels relies on a specially developed computer platform providing the sound pressure wave, the F0 contour, the segmental labels and the acoustic output (for further details see [13.,16.]). So far only 1/3 of the labelled spontaneous data have prosodic markers.

5. DATA BANK SEARCH AND SPEECH PROCESSING

The label files contain all the relevant phonetic information about spoken texts:

- corpus references
- orthographic forms of words and non-lexical items
- canonical transcription transforms, including word boundaries and function word markers
- segmental and componential labels with their time marks
- time-aligned utterance prosody labels.

These symbolic label files, in conjunction with the acoustic data base of read and spontaneous speech files, are prerequisite to setting up an annotated phonetic data bank, which, together with incorporated retrieval tools, allows corpus searches and speech processing of referenced signals for any phonetic research question in the phonological and acoustic domains.

5.1. The Canonical Lexicon

Thus a canonical pronunciation lexicon can be generated for a corpus, listing all the word types in orthographic and in canonical form, together with their frequencies of occurrence (see **Table 1**) [3.,4.,5.,7.].

Anschluß	Q'an#Sl"Us	4
April	Qapr'Il	11
Arbeit	Q'a6baIt	3
Arbeiten	Q'a6baIt@n	1
Arbeitsfrühstück	Q'a6baIts#fr"y:#St"Yk	5
Arbeitsfrühstücks	Q'a6baIts#fr"y:#St"Yks	1
Arbeitskreis	Q'a6baIts#kr"aIs	1
Arbeitssitzung	Q'a6baIts#z"ItsUN	2
Arbeitssitzungen	Q'a6baIts#z"ItsUN@n	2
Arbeitstreffen	Q'a6baIts#tr"Ef@n	8

Table 2: Excerpt from the canonical lexicon for CD-ROM#2 of the 'Kiel Corpus of Spontaneous Speech'.

5.2. Variants Lexicon

Similarly, a variants lexicon can be derived for a data base, containing orthographic, canonical and label forms for each entry as well as frequencies of each lexical item and labelling, together with the corpus references (dialogue session, turn and serial word number within the turn; see **Table 3**) [3.,4.,5.,7.].

Arbeit	Q'a6baIt	Q-:q'a6baIt	3	1	G085A006	22
Arbeit	Q'a6baIt	Q-:q'a6baIt-:	3	1	G097A000	17
Arbeit	Q'a6baIt	Q=-z:'a6baIt-:	3	1	G085A008	13
Arbeiten	Q'a6baIt@n	Q-:q'a6baIt-q@-:n	1	1	G096A000	19

Table 3: Excerpt from the variants lexicon for CD-ROM#2 of the 'Kiel Corpus of Spontaneous Speech'.

5.3. Data Analyses

The Kiel data bank has also been used to carry out the following language and speech investigations in German:

- glottal stops and glottalization for word-initial vowel onset and as plosive realization [8.,12.,15.,17.]
- vowel deletion [2.]
- realization of schwa syllables [11.,15.]
- phonetic variants of function words [20.]
- duration of stressed vowels [10.].
- acoustic analysis of hesitation particles [19.].

6. CONCLUSION

The structural frame for a computer data bank of spoken German and its integration into basic and applied phonetic research has been developed at IPDS Kiel and is continually being filled with segmentally and prosodically labelled data from read and spontaneous speech recordings. In future more diverse types of spontaneous interactions will be included in the Kiel Corpus. The range of phonetic analyses will be extended, also to include prosodic variables and comparisons of speaking styles. Data bank analyses will provide rules for connected speech processes, which in turn will assist in the development of automatic segmentation and labelling procedures. Finally, this phonetic data bank concept can be extended to other languages to initiate large-scale multilingual phonetic comparisons.

7. REFERENCES

AIPUK = Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel

1. Carlson, R., Granström, B., Hunnicut, S. "Multilingual text-to-speech development and applications," in W.A. Ainsworth (ed), *Advances in Speech, Hearing, and Language Processing*, pp. 269-296, London: JAI Press, 1990.

2. Helgason, P., and Kohler, K.J. "Vowel deletion in the Kiel Corpus of Spontaneous Speech," *AIPUK* 30: 115-158, 1996.

3. IPDS, *CD-ROM#1: The Kiel Corpus of Read Speech, vol. I*, Kiel, IPDS, 1994.

4. IPDS, *CD-ROM#2: The Kiel Corpus of Spontaneous Speech, vol. I*, Kiel, IPDS, 1995.

5. IPDS, *CD-ROM#3: The Kiel Corpus of Spontaneous Speech, vol. II*, Kiel, IPDS, 1996.

6. Karger, R., and Wahlster, W., *VERBMOBIL Handbuch, V3, Verbmobil Technisches Dokument 35*, DFKI, Saarbrücken, 1995.

7. Kohler, K.J., Lexica of the Kiel PHONDAT Corpus: Read Speech, vols. I,II, *AIPUK* 27 & 28, 1994.

8. Kohler, K.J. "Glottal stops and glottalization in German. Data and theory of connected speech processes," *Phonetica* 51: 38-51, 1994.

9. Kohler, K.J. "Complementary phonology: a theoretical frame for labelling an acoustic data base of dialogues", *Proc. ICSLP94* 1: 427-430, 1994.

10. Kohler, K.J. "Stressed vowel duration in German," *Acta Linguistica Hafniensia* 27: 299-321, 1994.

11. Kohler, K.J. "Articulatory reduction in different speaking styles," *Proc. XIIIth ICPhS* 2: 12-19, Stockholm, 1995.

12. Kohler, K.J. "The realization of plosives in nasal/lateral environments in spontaneous speech in German," *Proc. XIIIth ICPhS* 2: 210 - 213, Stockholm, 1995.

13. Kohler, K.J. "PROLAB - the Kiel system of prosodic labelling," *Proc. XIIIth ICPhS* 3: 162-165, Stockholm, 1995.

14. Kohler, K.J. "Parametric control of prosodic variables by symbolic input in TTS synthesis," in J.P.H. van Santen, R.W. Sproat, J.P. Olive, J. Hirschberg (eds.), *Progress in Speech Synthesis*, Berlin/Heidelberg/New York/Tokyo, Springer-Verlag, 1996.

15. Kohler, K.J. "Phonetic realization of German /ə/-syllables," *AIPUK* 30: 159-194, 1996.

16. Kohler, K.J., Pätzold, M., Simpson, A.P., From Scenario to Segment: the Controlled Elicitation, Transcription, Segmentation and Labelling of Spontaneous Speech, *AIPUK* 29, 1995.

17. Kohler, K.J., and Rehor, C. "Glottalization across word and syllable boundaries," *AIPUK* 30: 195-206, 1996.

18. Pätzold, M., Scheffers, M., Simpson, A.P., Thon, W. "Controlled elicitation and processing of spontaneous speech in Verbmobil," *Proc. XIIIth ICPhS* 3: 314-317, Stockholm, 1995.

19. Pätzold, M., Simpson, A.P. "An acoustic analysis of hesitation particles in German," *Proc. XIIIth ICPhS* 3: 512-515, Stockholm, 1995.

20. Rehor, C. "Phonetische Realisierung von Funktionswörtern im Deutschen," *AIPUK* 30: 1-114, 1996.

21. Tillmann, H.G., Draxler, Ch., Kotten, K., Schiel, F. "The phonetic goals of the new Bavarian Archive for Speech Signals," *Proc. XIIIth ICPhS* 4: 550-553, Stockholm, 1995.

22. Thon, W., and van Dommelen, W. "PHONDAT 90: Rechnerverarbeitbare Sprachaufnahmen eines umfangreichen Korpus des Deutschen," *AIPUK* 26, 41-79, 1992.

23. Thon, W. "Struktur eines Datenverarbeitungssystems für das Kieler PHONDAT-Projekt: Von der Aufnahme ASL-PHONDAT 92 zur Datenanalyse," *AIPUK* 26, 111-173, 1992.

24. Duden, *Der Duden Bd. 1, Rechtschreibung der deutschen Sprache*, 20. Aufl., Dudenverlag, Mannheim/Wien/Zürich, 1991

ACKNOWLEDGEMENT

Part of the work reported here was carried out with financial support from the German Ministry of Education, Science, Research and Technology (BMBF) under VERBMOBIL contract 01IV101M7.