

The Influence of scenario constraints on the spontaneity of speech

A comparison of dialogue corpora

Karl Weilhammer*, Daniela Oppermann*, Susanne Burger†

*Institute of Phonetics and Speech Communication
Schellingstr. 3, 80799 Munich, Germany
{karl.weilhammer, daniela.oppermann}@phonetik.uni-muenchen.de

†Interactive Systems Laboratories
Carnegie Mellon University Pittsburgh, USA, University of Karlsruhe, Germany
sburger@cs.cmu.edu

Abstract

In this article we compare two large scale dialogue corpora recorded in different settings. The main differences are unrestricted turn-taking vs. push-to-talk button and complex vs. simple negotiation task. In our investigation we found that vocabulary, durations of turns, words and sounds as well as prosodical features are influenced by differences in the setting.

1. Introduction

Spontaneous speech is one of today's great challenges for phonetics, linguistics and speech technology. Especially for the development of speech recognisers large corpora have been recorded and transliterated. In this context a distinction between spontaneous speech and read speech is well established. But after training a recogniser with one type of spontaneous speech it might fail to correctly recognise spontaneous speech of a different corpus even when the language model is adapted.

2. The corpus

During the VERBMOBIL project a great number of spontaneous speech dialogues have been recorded and transliterated (Wahlster, 1997). In all the recordings two German native speakers were asked to fix a date for a business meeting (Kohler et al., 1994) by using previously prepared diaries. For these recordings transcriptions of orthography and concomitant phenomena have been carefully prepared by humans (Burger, 1997). The obtained Corpus can be divided into two parts.

2.1. Push-to-talk buttons and simple task

We took 789 spontaneous speech dialogues from the corpus that was prepared during the first part of the Verbmobil. These are 317142 words occurring in 13910 different utterances (turns).

The scenario was modified slightly in different ways. In the simplest case subjects were asked to fix up to three dates on a business trip or for a short meeting. More complicated tasks included arranging a flight or a train to their appointment or were even complete travel planning. The dialogues were recorded in Kiel, Bonn, Karlsruhe and Munich, which means that the vocabulary of each site is influenced by different regional variants of standard German.

In most of the cases the subjects had to press a push-to-talk button to start the recording of their utterances. This means that only the speaker that has pressed the button first is recorded and can not be interrupted by his vis-à-vis until he releases the button. In this situation turn-taking is

ruled by pressing the button and not by prosodic, syntactic or gestural keys.

2.2. Unrestricted turn-taking and complex task

In the second phase of the project the scenario was more complex. The subjects were asked to fix a date for a one and a half day business trip. Additionally they should decide which plane or train they want to take and in which hotel they want to stay. Finally they have a short conversation about the evening programme. In most of the dialogues both persons were in the same role. For one seventh of the dialogues one subject should request information on the above described topics and the other should supply him/her¹ with the details.

The subjects were seated face to face on a table. Turn-taking followed more or less the rules of a natural conversation, even though the subjects were told not to interfere each other, but to wait until the person speaking has finished his turn. During the whole dialogue each subject was recorded on a separate channel and based on the recording every utterance was carefully transliterated.

For this investigation we picked 205 dialogues consisting out of 112513 uttered words that were produced in 8473 utterances.

3. Vocabulary

After 113000 uttered words the vocabulary of the complex task has a size of 4367 words compared to 3909 for the simple scenario. Figure 1 shows that in the beginning both curves are almost identical. After about 10000 uttered words the most frequent vocables which occur in almost every text or utterance have been included into the vocabulary. From this point the influence of the different tasks on the vocabulary size increases. Obviously the vocabulary rise for the complex task is higher, because the subjects talk about several different topics and will therefore introduce more new words (Weilhammer and Burger, 1998).

¹To keep the text easy to comprehend we will only use the male form, although the subjects were male and female.

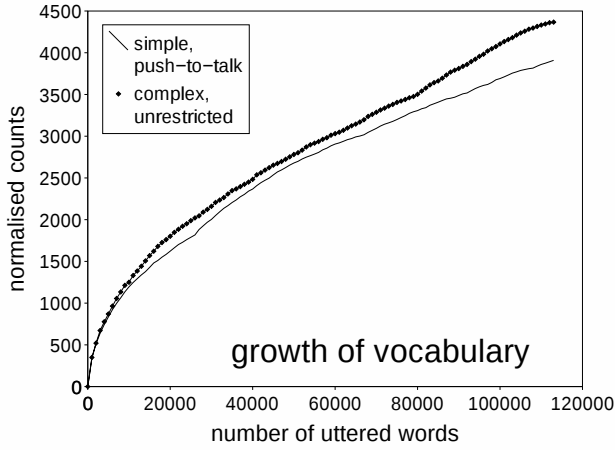


Figure 1: Increase of vocabulary with the number of uttered words for the simple and the complex conversation task.

4. Turn length

We measured the turn length in words, sounds and in seconds. The turn length in words was derived from the transcripts prepared by human transcribers. The values for the turn length in sounds are obtained from a SAMPA phoneme transcription that was automatically generated by MAUS (Munich AUtomatic Transcription System) (Kipp et al., 1997). The calculations for the turn durations are based on manual turn segmentations, from which we subtracted the length of initial and final silence segments that were also automatically generated by MAUS. In Figure 2 frequency distributions of the turn length in words and sounds are plotted together with a histogram of the turn duration (Width of bins 1 second). All graphs are normalised with regard to the number of turns that were examined respectively.

4.1. Very short turns

The number of very short turns in the dialogues with unrestricted turn-taking is in all three plots much higher than in the dialogues with push-to-talk button. These mini turns occur in a natural dialogue because the listener signals the speaker now and then, that he can follow his words by saying eg. “yes” or “mhm”. The positions of these “affirmative interjections” and all other interferences are marked in the transliterations. They will be subject to further research.

4.2. Turn length in words

With push-to-talk button the number of turns of a length in the range of 1 to 20 words remains quite constant and decreases then slowly. In the case of unrestricted turn-taking the number of turns of a certain length decreases slower after the “mini-turn peak”. It crosses the push-to-talk-button line at about 13 words per turn and remains below it. We don’t have enough data to verify a further crossing point, for a turn length of more than 100 words.

4.3. Turn length in sounds

After the “mini-turn peak” the line representing natural turn-taking decreases from a very low level very slowly.

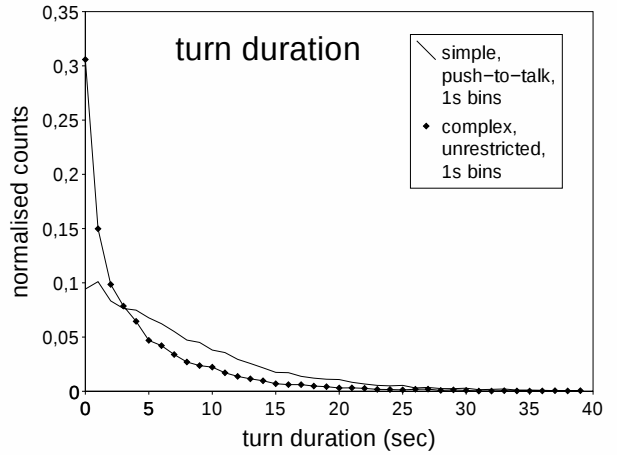
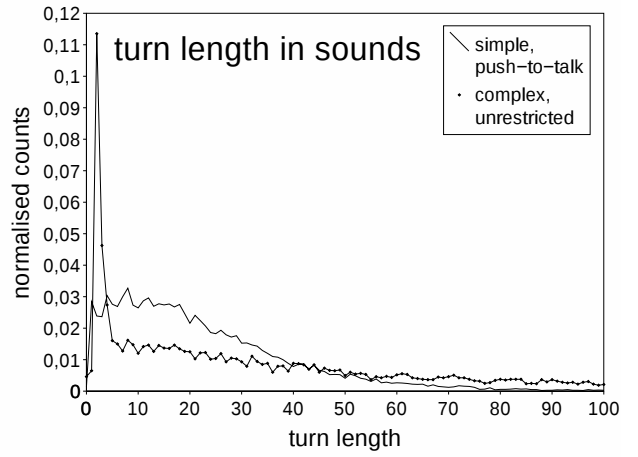
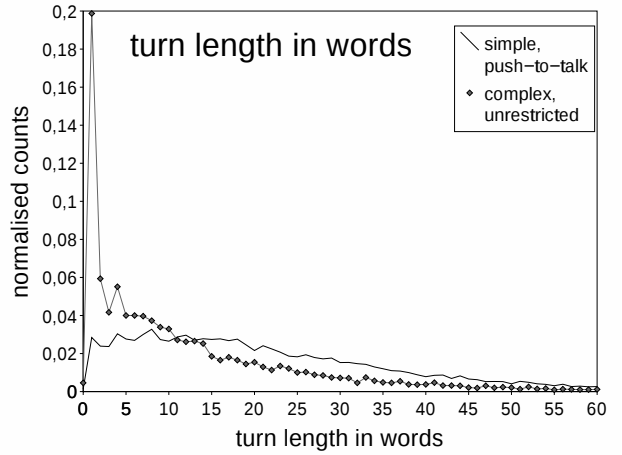


Figure 2: Influence of turn-taking restrictions on the turn length.

In the push-to-talk-button setting turns with 4 to 20 sounds are almost twice as frequent as turns with the corresponding length in the unrestricted case. Turns with around 40 sounds are equally likely in both scenarios. Turns longer than that are more often found in the case of unrestricted turn-taking.

4.4. Turn duration in seconds

Turn duration in seconds shows a similar behaviour than turn length in words. Both curves have almost perfect hy-

perbolic shape. Turn longer than 3 milliseconds are more frequent in the case of artificial turn-taking using the button. As in 4.2. we don't have enough data to verify a further crossing point, for a turn duration of more than 50 seconds.

4.5. Discussion of turn length

The examination of the turn length showed that natural dialogues tend to have a lot of very short turns which rather control the process of information exchange than transmit information related to the subject-matter. Turns shorter than 4 seconds with a length in between 4 and 15 words are more frequent in unrestricted turn-taking. Longer turns are more likely to occur in recordings with push-to-talk button. This suggests that the concept of the push-to-talk button protects the speaker from being interrupted and therefore causes longer turns.

Since the "short-turn" peak curve representing the number of sounds per turn covers much more area than the corresponding peak for words, there seem to be effects that have to do with the accuracy of articulation or the word length within turns.

5. Word length

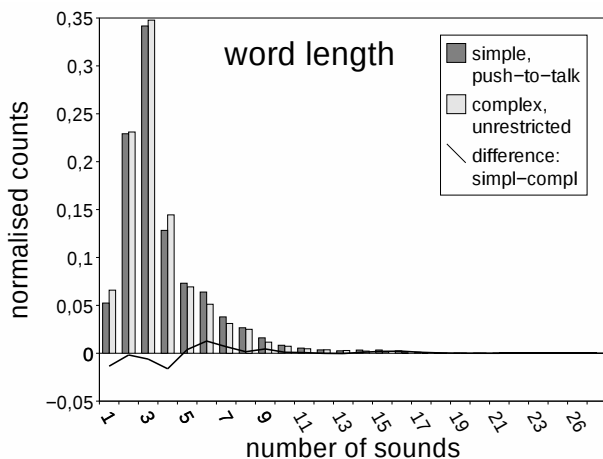


Figure 3: Distributions of word length in sounds for the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

The word length for all uttered words was measured in sounds and in seconds. The values for the word length in sounds and the sound durations are obtained from a SAMPA phoneme transcription and the corresponding phoneme segmentation that were both automatically generated by MAUS. Figure 3 shows the word length in sounds as they were actually uttered in the recordings and Figure 4 gives the word length in sounds as they should be uttered in their well-defined canonical form. Figure 5 shows a histogram of word duration (Width of bins 1 second). All graphs are normalised with regard to the number of words that were examined respectively.

5.1. Word length in sounds

Figure 3 shows that words with less than five sounds are more frequent in the case of the more complex sce-

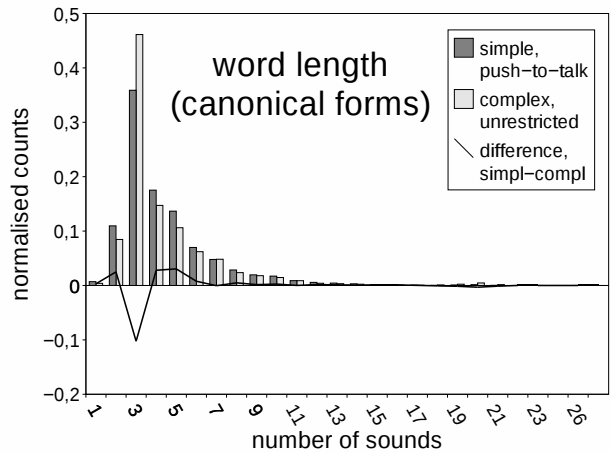


Figure 4: Distributions of word length in sounds for the canonical forms of all words in the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

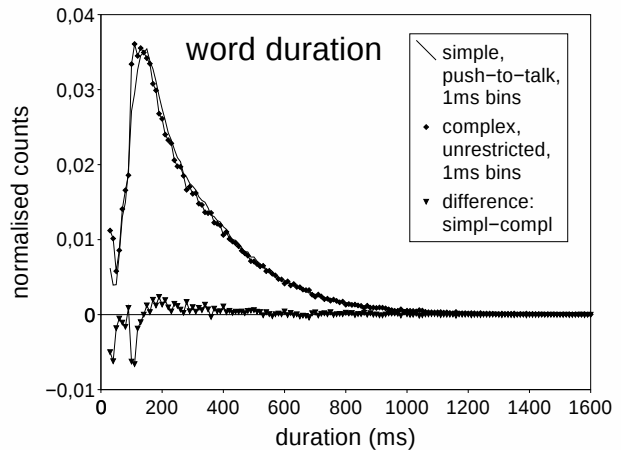


Figure 5: Histogram of word duration with 1 msec bins. For the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

nario with unrestricted turn-taking than in the simple task with push-to-talk button. For the canonical forms (Figure 4) we find this behaviour only for three-phoneme words. Since the canonical form is always longer than the corresponding actual utterance it is very likely that many of the three-phoneme words were reduced to one- or two-phoneme words in spontaneous speech. The relatively high number of four-phoneme words comes from reductions of canonical forms of longer words.

5.2. Word duration

The curves of the word-duration histograms in Figure 5 are shaped similarly. It almost looks as if the curve representing unrestricted turn-taking was shifted a little to shorter durations compared to the push-to-talk-button curve. This effect explains the two negative peaks in the difference of the two distributions at 40 and 100 ms. For

words longer than 1300 ms the difference becomes negative again, even though this is not very reliable, because the counts are very low. .

5.3. Discussion of word length

For the unrestricted case the distributions of the Figures 3 and 5 show an increase of short words that is caused by different reasons: A different degree of phoneme reductions in spontaneous speech in both corpora is one of them. The contribution of frequency differences of words and different vocabulary entries can be obtained from Figure 4.

6. Sound duration

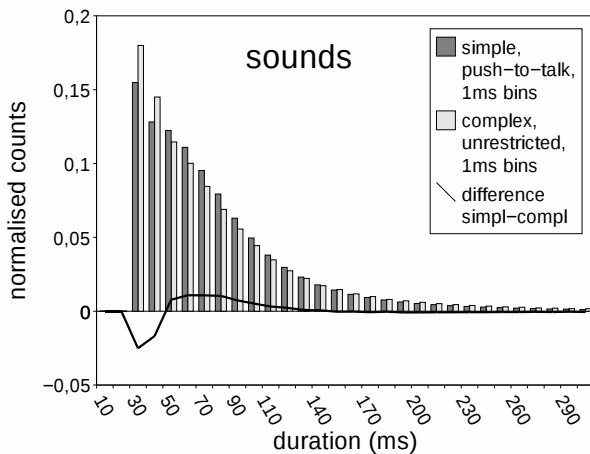


Figure 6: Histogram of sound duration with 1 msec bins. For the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

Our investigation of sound length is based on the automatic segmentation system MAUS. Using these data one must be aware that MAUS tends to produce shorter segments than human transcribers. But using the same system on two spontaneous speech dialogue corpora should at least give some hints whether there are differences in sound length. Figure 6 shows the distributions of sound length for both tasks and their difference. For the complex task we find more very short sounds and sounds longer than 150ms. In the medium range the distribution of the simple push-to-talk task has higher values. To get a more distinct picture we prepared histograms for different sound categories.

In the unrestricted turn-taking task there is only one major difference, which is a peak at the number of short consonants that is much higher than in the push-to-talk-button task.

We produced different distributions for long and short vowels (Long and short in a phonological sense). Comparing the distributions for long and short vowels we find that the distribution for the long vowel is indeed wider than the corresponding one for the short vowel.

In the unrestricted scenario short vowels (Figure 7) with a duration of less than 100ms were more frequent than the corresponding short vowels in the push-to-talk button task.

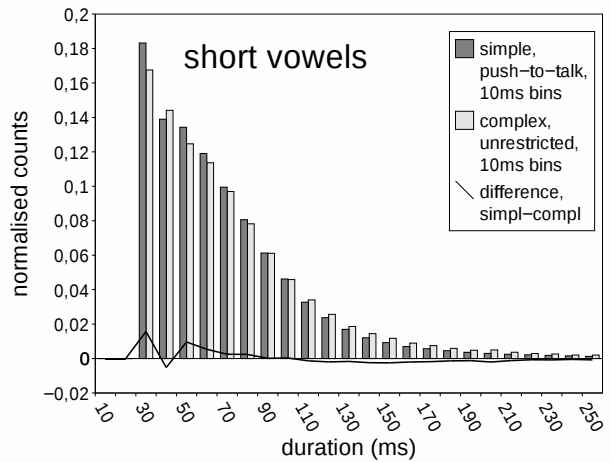


Figure 7: Histogram of durations of short vowels with 1 msec bins. For the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

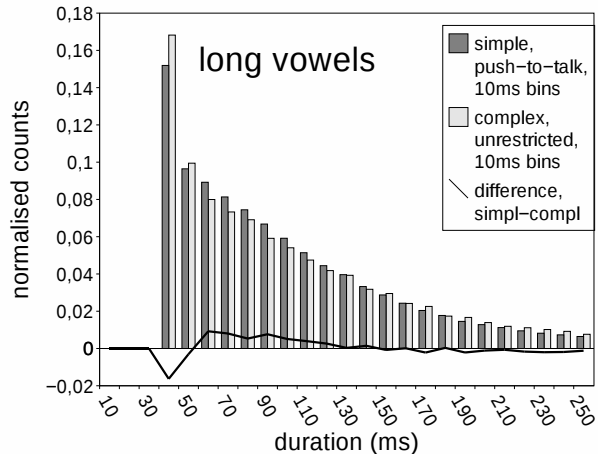


Figure 8: Histogram of durations of long vowels with 1 msec bins. For the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

For long vowels and consonants we observe a behaviour similar to sounds.

This means that long vowels and consonants have a greater variance in duration in the case of natural turn-taking, while short vowels tend to have longer durations.

7. Prosodic features

7.1. Pauses in speech

Speech pauses are used to impose some structure on text. Figure 10 shows the distribution of the duration of pauses in both corpora. When the subjects could use natural turn-taking they often used pauses of less than 50ms. With the push to talk button they used long pauses more frequently. The peaks at 10 and 30ms may be artificial, because usually there is a silence interval of this duration before and after speech starts in each the recording, which

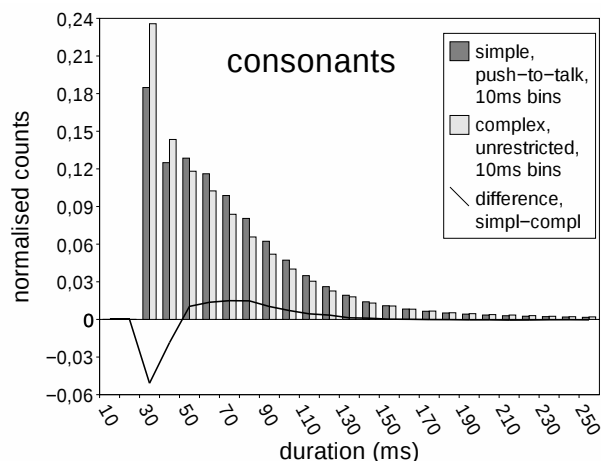


Figure 9: Histogram of durations of consonants with 1 msec bins. For the recordings with push-to-talk button and with natural turn-taking, furthermore the difference of these distributions.

unfortunately were counted as well.

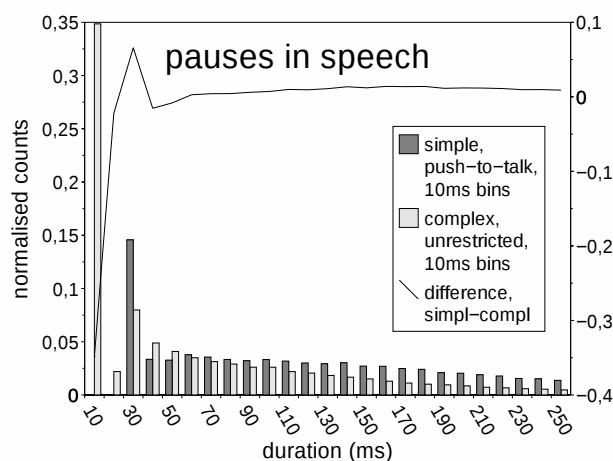


Figure 10: Histogram of durations of pauses in speech with 1 msec bins. For the recordings with push-to-talk button and with natural turn-taking (Right axis). The difference of these distributions is plotted with respect to the left axis.

7.2. Fundamental frequency

We calculated the fundamental frequency over the recorded data of the male speakers of each corpus using a Schaefer-Vincent periodicity detector and created a histogram (Figure 11). The baseline values are 119Hz for the subjects that were using a push-to-talk button and 107Hz for unrestricted turn-taking. According to (Cruttenden, 1997) speakers tend to rise their voice in a stress situation. The increase of fundamental frequency in the push-to-talk task could mirror the exertion of the subjects who feel forced to speak after having pressed the button.

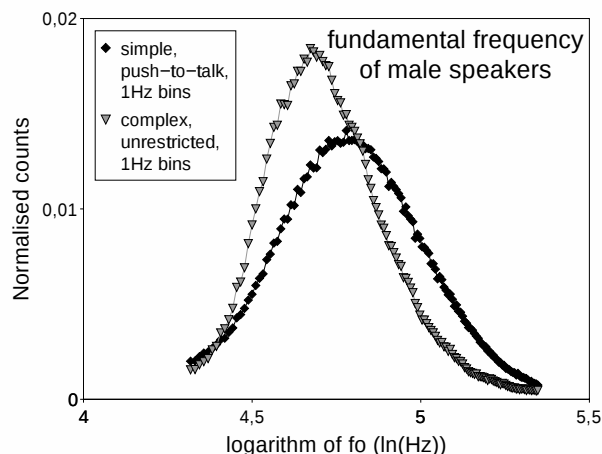


Figure 11: Histogram of fundamental frequency of male speakers with 1Hz bins. The counts for the histogram were made on a linear scale. For the recordings with push-to-talk button and with natural turn-taking.

8. Conclusion

We could show many examples for differences in the quality of speech invoked in a natural turn-taking situation compared to regiment turn-taking by using a push-to-talk button. In the artificial situation one loses very short turns that control turn-taking, and gets longer speech pauses and longer words. The variance of the sound duration is smaller. Caused by the unnatural situation the fundamental frequency rises. Therefore it is necessary to introduce further distinctions when characterising spontaneous speech.

9. References

- Burger, Susanne, 1997. Technical report, Transliteration spontansprachlicher Daten, Lexikon der Transliterationskonventionen in Verbmobil II. Verbmobil TechDok-56-97, available via Internet: <http://www.phonetik.uni-muenchen.de/VMTechDoks.html>.
- Cruttenden, Alan, 1997. *Intonation*. Cambridge University Press, 2nd edition.
- Kipp, Andreas, Barbara Wesenick, and Florian Schiel, 1997. Pronunciation modeling applied to automatic segmentation of spontaneous speech. In *Proceedings of the EUROSPEECH 1997, Rhodes, Greece*.
- Kohler, Klaus, Gloria Lex, Matthias Pätzold, Michael Scheffers, Adrian Simpson, and Werner Thon, 1994. Technical report, Handbuch zur Datenerhebung und Transliteration in TP14 von Verbmobil-3-0, Verbmobil Techdok-11-94.
- Wahlster, Wolfgang, 1997. Technical report, Verbmobil: Übersetzung von Verhandlungsdialogen, Verbmobil report-01-93, available via Internet: <http://www.dfki.de/cgi-bin7verbmobil/htbin/doc-access.cgi>.
- Weilhammer, Karl and Susanne Burger, 1998. Characterising a database of spoken German by techniques of data mining. In *Proceedings of the First Conference on Language Resources, Granada, Spain*.