WHAT MAKES SPEECH DATA SPONTANEOUS?

Daniela Oppermann[1], Susanne Burger [2]

[1]*Institute of Phonetics and Speech Communication, University of Munich, Munich, Germany,*
[2]*Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, USA,*
University of Karlsruhe, Karlsruhe, Germany
[daniela@phonetik.uni-muenchen.de][sburger@cs.cmu.edu]

ABSTRACT

The aim of the work to be reported here is the development of schemata which are able to predict the quality of spontaneity and help to create and collect databases for certain tasks on spontaneous speech. The term "spontaneous speech" is used in a wide range and allows the existence of many spontaneous speech corpora with different levels of spontaneity. Our aim was to find appropriate categories and description values for these corpora. Therefore we started by analyzing transcriptions of spontaneous monologues of one minute, which were recorded and annotated. We made a structure analysis of the introductory part of the monologues and let people qualify categories of spontaneity in a small experiment containing a subset of the monologues. Correlations between the judged categories of spontaneity and the amount of spontaneous speech phenomena in the monologues will be shown.

1. INTRODUCTION

In recent years speech recognition has concentrated more and more on understanding spontaneous speech with the aim that every person should be able to communicate with speech understanding systems in natural language. To train these systems it is necessary to create corpora which are based on spontaneous speech data. There already exist various types of speech corpora all of which could be considered as spontaneously spoken language. The problem is that all these databases refer to varying levels of spontaneity. In fact there is a wide range in the levels of spontaneity due to different kinds of requirements of recognition systems. Several factors can influence the level of spontaneity, e.g. the required amount of the vocabulary, the domain, or the microphone used. Another important feature is the question whether the data is based on telephone recordings and what kind of instructions and tasks were given to the speakers. All these conditions led to the development of many different kinds of speech corpora, which are supposed to contain spontaneous speech data, but are not comparable and not compatible to each other or could hardly be repeated or complemented.

The current search of factors, which should allow predicates about the level of spontaneity of collected speech data, is part of a serial of experiments we are realizing on the approach of improving our collections and searching of evaluation sets for existent databases [4][5].

The focus of the present work lies in the development of categories which are able to describe what elements of speech allow people to decide whether speech is spontaneous or a prepared talk and how is the structure of spontaneous speech. We tried to establish this by means of an analysis of the RVG1 corpus [1] which is a regionally covered collection of aspects of currently spoken German.

The first experiment describes what we found within our material regarding the internal speech act structure. Further we tried to get appropriate judgments from subjects about the level of spontaneity of presented texts and finally we searched for correlations between the so called spontaneous phenomena like breathing or hesitations, etc. and the judgments we got from our subjects.

The resulting schemata could be used in comparing spontaneous speech corpora and should be taken into account before new databases in this field are collected.

2. DATABASE

The database for this study consists of transliterated recordings of spontaneous monologues which are part of the RVG1 corpus (Regional Variants of German) [2]. This corpus was recently recorded at the Phonetics Department at Munich University in co-operation with AT&T, Lucent Technologies and the Bavarian Archive for Speech Signals (BAS) [9]. It covers all regional variants of German, including the German dialects spoken in Switzerland, Austria, and Northern Italy. With regard to the main task of collecting currently spoken German, the determination of how many speakers of each German-speaking region are to be recorded was made by means of population density and according to the dialectal subdivision introduced in [6]. All RVG1 recordings were made in a quiet room. In total, the RVG1 corpus consists of 42500 read utterances (polyphone-type material: single digits, digit sequences, commands, phonemically rich sentences, telephone numbers) and 491 spontaneous monologues spoken by 491 speakers (43% female, 57% male). Every spontaneous monologue lasted one minute. The speakers got their explanations by prompts on a PC screen. They didn't get any specific order about the topic of their talk.

The recording situation was as follows: first the speakers had to read aloud the digits, commands and various sentences which were promted on a PC screen. This task lasted about 20 minutes. At the end of this session the subject had the task, also given via PC screen, to narrate something spontaneously. A lot of speakers were surprised when confronted with this situation, so we added some proposals like "what did you do today" to the displayed explanation. The monologues were transliterated on the

orthographic level according to Duden [8]. The transliteration conventions follow the standard for the transliteration of spontaneous speech as defined in VERBMOBIL [2]. Attention was also directed to the annotation of pronunciation variants. A striking deviation from the standard pronunciation was annotated by means of additional comments appended to the standard orthographic transliteration of the word concerned. A more detailed description of the rules used for the annotation of pronunciation variants can be found in [3].The transcribers are trained students of German Linguistics. There was more than one person concerned with the transliteration of each monologue. This was to reduce the influence of a transriber's dialectal origin.

3. ANALYSES

3.1 Structure analysis

3.1.1 Introductory part. A first next step of the study was to categorize all 491 monologues according to their internal speech act structure. We read the monologues and categorized their introductory material, which lasted in general one or two sentences, according to their internal structure.

At the very beginning, the first term speakers started with could be divided into six broad categories. We categorized the beginning with "interjection" when the monologue started with "well" or "okay", or something similar. The category "hesitation" stands for the annotation of hesitant phenomena like "uhm" or "uh". Some speakers started their monologue with a question like "may I start now?" and others simulated dialogues and began their talk addressing a person, which we also counted as a category. People often started their talks giving a time or date to which we refer here as the category "date". The category "topic" was chosen when the speakers immediately went into the topic which indicated at the same time the beginning of the monologue body.

In Figure 1 there are given the frequency in percent of the categories which appeared in the first introductory part of the monologues.

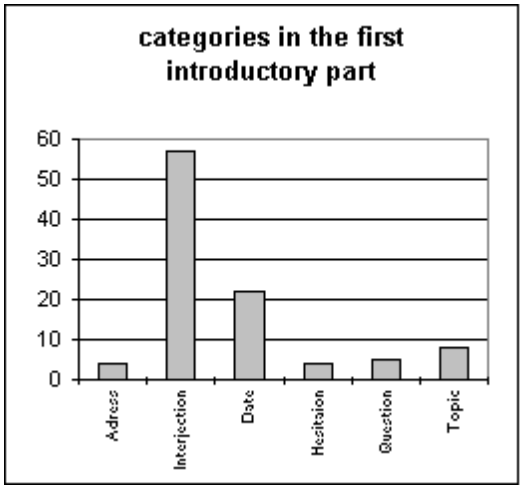

Figure 1: frequency of the categories in the first introductory part of the monologue (given in percent)

As a first result it should be mentioned that only 8% of the subjects knew immediately what they wanted to talk about (which is indicated by the category "topic").

Further, more than half of the monologues start with interjections. Obviously people tend to use "introduction features" when they are asked to narrate something without any further limitations or instructions. These "interjections" are probably a filler for bridging the gap between thinking and talking.

In 92% of all the cases in which people don't start immediately into the topic, further introductory structures can be observed in this part. Only four of the six previous categories appeared, while most of them belong to the category "date" (52%). In a few cases people used "interjection" (0,5%) or "address" (1,5%), as can be seen in Figure 2. The categories "question" and "hesitation" disappeared completely.

Almost half of the speakers (46%) led up now to the main body of the monologue which is indicated again by the category "topic".

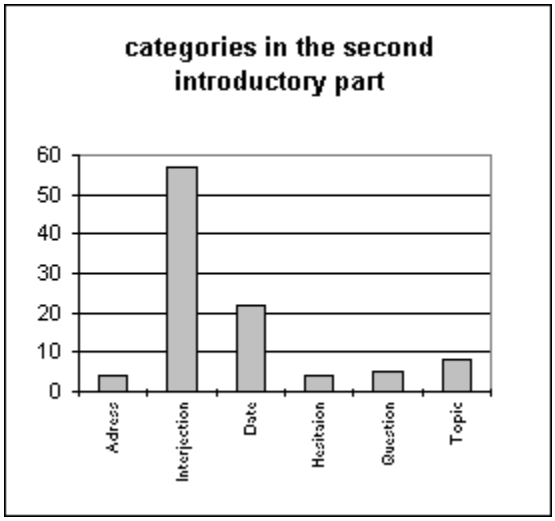

Figure 2: appearance of the categories in the second introductory part of the monologue (given in percent)

After this second introductory part all speakers without any exceptions made the transition to the monologue body and into their topic, without any further filling features.

This structure analysis revealed that most people begin with an interjection followed by a date or time when they are confronted with the task of narrating something for one minute, without any further instructions.

3.1.2 Change of topics. A further step in this investigation was to look at how often people switch to another topic within the one minute of monologues.

The results of this analysis are given here in percent.

6 or more changes	0%
5 changes	1%
4 changes	3%
3 changes	4%
2 changes	16%
1 change	18%
no change	59%

Table 2: change of topics within the monologues in percent

From Table 2 it can be seen that none of the speakers had a change in topic more than 5 times in the whole corpus, and also a change of three or four topics didn't appear very often (4%, 3%) in our material. In 16% of the monologues people switched to another topic twice and only once in 18% of the cases. But most of the speakers (59%) stuck to just one topic throughout the whole recording.

3.2. Spontaneity judgments

For our further analysis of the RVG1 corpus we made an experiment to get judgments of subjects about the spontaneity of the monologues. We decided to start with asking for the impression the subject get from simply reading the transliterations of a subset of the 491 monologues. We choose two monologues from each of the 24 regions (see map in [7]), which was about 10% of the whole corpus, and presented them to two groups of 12 test persons each. In order to make the text easily readable for persons who are not familiar with the annotation system, we changed the formats of the original transliterations. The layout of the resulting texts were similar to an actor's script.

In other words, we placed the annotations of pauses and breathing always in the beginning of a line, separated by tab for to simulate a break in fluency. We left in the text annotations of hesitations and lengthening of sounds which were indicated by underscores added to the concerning sound. In the case of commented pronunciations we substituted the pronunciation comment for the orthographic version, so that the reader had a better impression of how the speakers of the transliterated texts had spoken in real life. Another reason for fitting the pronunciation comments directly into the text, is that this would facilitate recognition of the regional origin of the speaker.

Here is an example of how the text was presented to the reader:

```
<Pause>
<Atmen>   also , eingklich hab' ich auch nix richtiges
          gemacht ,
<Atmen>   "ahm so f"u' die Uni oder so , son'ern ich hab'
          pfff in erster Linie mein Zimmer aufger"aumt , un'
<Atmen>   wir kriegen 'ne neue Mitbewohnerin , und na
          mu"sten wa dat Zimmer noch renovieren
<Atmen>   un' dat is' ziem'ich viel Arbeit , da m"uss' ma
          streichen un' so ,
```

The readers had to fill out :

- whether the originally talk was a spontaneous realization or whether it was prepared previously
- whether the text contains more fluent speech or more hesitant
- whether the sentence construction within the text makes a simple impression or if the sentences were complex
- whether the subjects used a simple vocabulary or a more complex vocabulary

As an additional task we asked the test persons to estimate by means of a printed map containing the Bundesländer of Germany, Switzerland and Austria which dialectal region the original speakers could be from.

The following Table 1 shows the decisions of the readers on the transliterated and specially formatted monologues.

spontaneous talk	94%
prepared talk	06%
fluent talk	58%
hesitant talk	42%
sentence construction simple	71%
sentence construction complex	29%
vocabulary simple	79%
vocabulary complex	21%
region recognized	46%
region unrecognized	54%

Table 1: judgments of the presented monologues in percent

The results of this investigation show that 94% of all the presented texts were judged as spontaneous. More than half of the texts (58%) were estimated as fluently spoken. 71% of the monologues had been categorized as spoken with simple sentences and even 79% with a simple vocabulary.

In 54% of all texts the readers were not able to recognize the regional affiliation of the speakers.

3.3 Correlation between spontaneity judgments and selected spontaneous speech phenomena

In a further step we wanted to know if there are any relations between the judgments of the persons who had to read the monologues and the number of annotated phenomena which occur in spontaneous speech as mentioned above. Examples of spontaneous phenomena are hesitations, breathing, pauses, lengthening of sounds, self corrections or repetition of words, etc.

We extracted these annotated phenomena automatically with the help of a parser and compared the number of occurrences with the judgments of the readers for every monologue.

The results are expressed in the following correlation matrix:

	spontaneous monologue	fluent talk	sentence simple	vocabulary simple	region recog.
hes	0,17	**0,29**	0,07	**0,26**	*-0,26*
zog	0,07	**0,4**	0	-0,15	0,11
agram	*-0,21*	0,17	**0,2**	-0,01	0
art_abr	*-0,43*	0,19	-0,08	-0,17	-0,17
atmen	-0,02	-0,07	0,14	0	0,19
pause	**0,28**	**0,5**	**0,36**	**0,29**	-0,14
proz_va	**0,33**	*-0,2*	**0,22**	**0,34**	**0,39**

Table 3: correlation matrix: x-axis: judgments of the readers, y-axis: extracted phenomena; hes = hesitations, zog = lengthening of sounds, agram = self corrections and word repetitions, art_abr

= break-off in the middle of words, atmen = breathing, pause = pause of more than 400 ms, proz_va = percent of pronunciation comments per total word amount of each monologue; postive correlated phenomena > 0,2 printed bold, negative correlations > 0,2 printed italic

From the correlation matrix it can be seen that there are some phenomena in spontaneous speech that might have influenced the readers in their choice of the appropriate category. The strongest influence is shown by the occurrence of pauses in the text, but an interesting feature is the positive correlation with the judgment of the fluency of talk. This could be interpreted as showing that the more pauses the speaker made the more the test person rated the text as fluent speech. This result could be explained by the layout used to present the transliterations to the readers. As mentioned above, the presentation of the texts was reminiscent of an actor's script; the pauses were separated from the text and might therefore be overlooked by the readers. The feature "pause" seems also to have influenced the ratings of the simplicity of the sentence construction and the vocabulary, as well as the ratings of levels of spontaneity. Another feature which had an effect on almost every decision was the phenomenon "hesitation". It influenced positively the ratings for the fluency of the text and the simplicity of the vocabulary, but it shows a negative effect on the regional recognition. The appearance of annotated hesitations in the text might have irritated the readers in recognizing the origin of the speakers. The feature "proz_va" seems to have quite strong effects on every category. At least for the recognition of the dialectal region this could have been expected. It has only a negative influence on the judgment of the fluency of speech. What is also a striking result is the negative influencing effect of repetitions, self corrections and broken-off words on the decision of spontaneity of the text. Here the question arises as to whether the influence on this category might be neglected, given that 96% of the monologues were judged as spontaneous.

The correlation between the judgments of the readers and the spontaneous speech phenomena becomes clearer if similar phenomena are grouped together into classes. Accordingly the categories "atmen" and "pause" were collapsed in to one group, as were the phenomena "hesitation" and "lengthening", and the phenomena "corrections" and "break-off" of words. At least within the first two classes the correlation with the judgment "fluency of talk" becomes stronger.

	sum_at_pau	sum_hes_zog	sum_agr_abr
fluency	0,32	0,45	0,19

Table 4: Correlations of spontaneous phenomena which were grouped together and the fluency of talk. sum_at_pau = summery of "atmen" and "pause", sum_hes_zog = summery of "hesitations" and "lengthening", sum_agr_abr = summery of "repetitions and corrections of words" and "words which were broken off"

It should also be mentioned that an influencing effect can also be found within the judgment categories themselves.

	sentence-simple	vocabulary-simple
fluency	0,47	0,37

Table 5: Correlation between the fluency of the talk and the simplicity of the sentences and vocabulary.

There seems to be a correlation effect not only between the spontaneous speech phenomena and the judgment categories, but also within these categories themselves. As can be seen in Table 5 the more people judged a text as fluently spoken, the more the sentences and the vocabulary of this text were considered simple.

4. CONCLUSION

In this investigation we could show that there exist specific spontaneous speech phenomena which can be analyzed and categorized. The analysis of the monologue structure revealed that people almost always started with some introductory features before going into the topic they wanted to talk about. The most frequent sequence of categories within the first one or two sentences was interjection before date and topic. We assume that these features could be a typical sign for this kind of speech corpus.

The high number of votes for the category spontaneous talk in the task where people had to judge the special formatted transliterations of the monologues confirmed our analysis concerning the spontaneity of the material. On the other hand the ratings may have falsified the correlation coefficients for the selected features. The influencing effects of the spontaneous speech phenomena are more prominent when they are related to the other judgment categories, especially when they were merged into larger classes.

The presented analysis is the beginning of further work on this kind of structure analysis. For future work we will try to find different categories and parameters which are able to reflect phenomena of spontaneous speech in a more appropriate way.

ACKNOWLEDGMENTS

This work couldn't have done without the experience and developments within the VERBMOBIL project [7].

REFERENCES

[1] Burger, S. 1998. RVG1 - A Prototype for the Collection of Current Spoken German. to be appear in *FIPKM.* München. 1999

[2] Burger, S. 1997. Transliteration spontansprachlicher Daten - Lexikon der Transliterationskonventionen - VERBMOBIL II. *Verbmobil TechDok-56-97.* München.

[3] Burger, S., Kachelrieß, E. 1996. Aussprachevarianten in der Verbmobil-Transliteration - Regeln zur konsistenteren Verschriftung. *VERBMOBIL Memo-111-96,* München.

[4] Burger, S., Oppermann, D. 1998. The Impact of Regional Variety upon Specific Word Categories in Spontaneous Speech. *Proceedings of ICSLP 1998,* Sydney

[5] Burger, S., Oppermann, D. 1999. Regional variants of German: Categories of pronunciation deviation from Standard German. to be appear in *Proceedings ICPhS.* San Francisco

[6] Burger, S., Schiel, F. 1998. RVG1 - A Database for Regional Variants of Contemporary German. *Proceedings of LREC 1998* Granada.

[7] Burger, S., Draxler, Ch. 1998. Identifying daialects of German from Digit Strings. *Proceedings of LREC 1998.* Granada

[8] Duden - Rechtschreibung der deutschen Sprache, *20. neu bearb. und erw. Aufl. Dudenverlag.* Mannheim, Wien, Züich. 1991.

[9] Schiel, F. 1997. The Bavarian Archive for Speech Signals: Resources for the Speech Community. *Eurospeech 1997.* (pp. 1687 - 1690).

[10] VERBMOBIL II: Verbmobil Homepage. *http://www.dfki.uni-sb.de/verbmobil*