THE IMPACT OF REGIONAL VARIETY UPON SPECIFIC WORD CATEGORIES IN SPONTANEOUS GERMAN

Susanne Burger & Daniela Oppermann

Department of Phonetics and Speech Communication, University of Munich, 80799 Munich, GERMANY
[burger|daniela@phonetik.uni-muenchen.de]

ABSTRACT

The aim of the work to be reported here is to verify that pronunciation variety in German spontaneous speech appears in specific linguistic word categories (POS) while others are less affected. Additionally, the material (transliterations of spontaneous monologues of the RVG1-corpus [5]) allows a detailed view on pronunciation regarding the speakers' origin within the German-speaking regions and regarding the topic a speaker is talking about. Our results show that generally the most affected parts of speech are the auxiliary verbs. The regions with the highest deviation rates of standard German are Switzerland and Austria, and the regions in southern Germany. We can only add the vague answer to the question, whether the semantically topic of a story may have an influence upon the deviation of the standard language, that our results show a striking less frequency of pronunciation variants, when people speak about their jobs.

1. INTRODUCTION

In recent years research in the field of speech recognition especially, has become interested in the impact that dialect or regional variation of a standard language can have upon recognition rates. Therefore, a major aspect in the collection of speech resources is the regional coverage of data within one language. Such recordings allow more detailed analysis of dialectal variants of the standard language, allow the training of recognition systems by means of appropriate speech data and can help in building regional sub-models of language models. A first experiment can be found in [1]. In this experiment the problem is discussed of whether moderate regional variants of German influence the automatic speech recognition process. The planning and creation of regionally covered databases requires great care regarding the aspects of how to get regional variants of a standard language. Also more information about the different levels of speaking styles, e.g. standard language, colloquial speech, slang, dialect., is needed. In the present work we concentrated the focus on the question how expressive the used annotation of the pronunciation variants may be. In this context several questions arise about

- the relation between the regional origin of a speaker and the number of annotated comments on pronunciation.
- the relation between the topic of the speakers' spontaneous monologue and the number of pronunciation comments.
- how specific linguistic word categories are affected by pronunciation comments

2. DATABASE

The database for this study are transliterated recordings of spontaneous monologues which are part of the RVG1 corpus (Regional Variants of German) [1]. 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) [10]. It covers all regional variants of German, including the German dialects spoken in Switzerland, Austria, and Northern Italy. All speakers of RVG1 were asked extensively about their regional background by an expert who classifies the speaker's accent according to well-defined dialect regions. With regard to the main task of collecting current 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 [9]. 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 500 spontaneous monologues spoken by 500 speakers (43% female, 57% male).

Every spontaneous monologue lasted one minute. The speakers got their directives by prompts on a PC screen and were asked to imagine talking to a person from the speaker's home region, but should avoid strong dialect. They didn't get any specific order about the topic of their talk. The monologues were transliterated on orthographic level according to Duden [7]. Special symbols for typical spontaneous phenomena like lengthening, hesitations, non-grammatical phrases and background noise are included. The transliteration conventions follow the standard for the transliteration of spontaneous speech as defined in VERBMOBIL [2]. Special attention was directed to the annotation of pronunciation variants. A striking deviation of the standard pronunciation was annotated by means of additional comments appended to the standard orthographic transliteration of the word concerned. Starting with a digit which indicates how many lexical elements are affected, the comment contains a written version of the diverging pronunciation in comment brackets. The representation of the variant remains as orthographic as possible, additionally, elisions are marked by apostrophes. The comments present a superficial impression of what the divergence looks like and gives an indication what lexical elements of a talk are affected.

The following examples show some typical pronunciation comments from the RVG1 monologues (the type of variation and the english translation is appended within the brackets):
haben wir <!2 hamma> (assimilation, we have)

nicht <!1 ned> (dialectal, not)
gesungen <!1 g'sungen> (reduction, sung*)*
kein <!1 keen> (dialectal, no*)*

The transcribers are trained students of German Linguistics. There is more than one person concerned with the transliteration of a monologue. This is for to reduce the influence of a transribers dialectal origin. A more detailed description of the rules used for the annotation of pronunciation variants can be found in [2,4,5].

3. MATERIAL

Currently, 413 of the monologues are prepared for further analyses and build the basis material for our studies. For the preparation of the experiment, we extracted only the lexical elements and the pronunciations comments from the transliterations. We counted the initial digits within the comments which denote the number of concerned lexical elements. The average percentage of variants per monologue is 23,6%, where 67% of the monologues show less variants as the average and 33% more than the average. Three monologues has the minimum variant frequency of 0%, two monologues appears at the maximum frequency at 77%.

code	category	class	freq
VAFIN	auxiliary finite verb	verb	19%
VAPP	auxiliary past participle verb		
VMFIN	modal finite verb		
VVFIN	content finite verb		
VVINF	content infinitive verb		
VVPP	content past participle verb		
PPER	irreflexive personal pronoun	pronoun	9%
PRF	reflexive personal pronoun		
APPR	preposition or left part of circumposition	prepos	8%
NE	proper noun	noun	14%
NN	common noun		
KON	coordinating conjunction	conjunc	8%
KOUS	subordinating conjunction with a sentence		
ART	definite or indefinite article	article	6%
ADV	adverb	adverb	18%
ADJA	attributive adjective	adjective	6%
ADJD	adverbially or predicatively used adjective		

Table 1: POS tag-set [8], code = abbreviation, category = linguistic term, class = broader POS classes, freq = total frequency in the entire corpus in percent

These texts were automatically POS-tagged [8]. The reference vocabulary for the tagging is a newspaper lexicon and the VERBMOBIL lexicon, based on transliterations of the VERBMOBIL spontaneous dialogues. Because of the unlimited scenario of the RVG1 monologues, about 24% of the tokens remained out of the used vocabularies and therefore, were tagged by hand. Further, we categorised the monologues with reference to their topics. The tagged material as well as the entire information we have about the speakers' dialect and origin, the filtered transliterations and the topic categorisations were calculated by means of Ms-Excel spreadsheets.

Table 1 shows the relevant parts of the POS tag-set. That means, we are presenting only the POS tags with interesting results concerning their variation comments. The row "code" shows the abbreviation according to [8], "category" gives the linguistic term and "class" contains the broad linguistic word classes. Due to the fact, that some POS categories appear only at a small number, we set the limit for relevance at an appearance frequency on at least 5%. The POS categories showed in table 1 represent 88% of the appearing POS categories. POS categories below 5% are not presented in table 1. The entire tag set can be found in [8].

4. RESULTS

4.1. Regions and pronunciation variants

The first part of our study shows which regions the speakers come from are to which extend affected by deviations from standard German. For this approach we counted the lexical elements per monologue and how often a lexical element had a pronunciation comment.

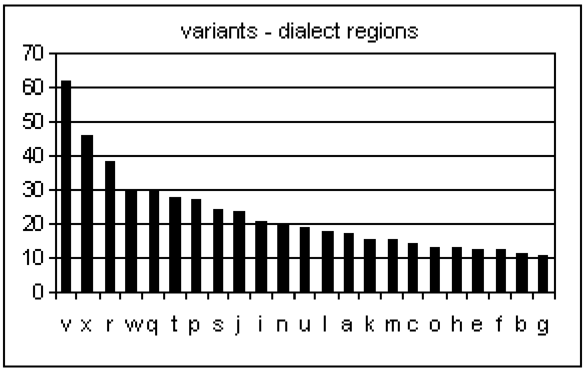

Figure 1: pronunciation comments in percent per monologue**,** dialect regions: a nordfriesisch, b ostfriesisch, c nordniedersächsisch, d mecklenburgvorpommersch, e ostfälisch, f westfälisch, g niederrheinisch, h mittelfränkisch, i moselfränkisch, j pfälzisch, k hessisch, l brandenburgisch, m thüringisch, n obersächsisch, o sorbisch, p ostfränkisch, q südfränkisch, r nordbairisch, s niederalemannisch, t schwäbisch, u mittelbairisch, v schweizerisch, w österreichisch, x tirolerisch [9]

As can be seen in figure 1, the highest number of deviations was found in Switzerland, where 62% of the lexical elements are commented upon and also in the north-western part of Austria, Tirol, with a 46% deviation from standard German. Together with the 30% commented words in the remaining part of Austria, it can be assumed, that the German language spoken outside of Germany tends to develop more individual characteristics. The results of the north Bavarian region "r" ("Oberpfalz") are also relatively high (38%). Together with the Franconian region "q" and the Swabian regions "t", "p" and "s" these monologues are the ones with a more than 20% deviation. The amount of pronunciation comments decreases the more the direction goes to the north-west of Germany. Actually, there is the well-known phenomenon, that the inhabitants of middle and north Germany avoid the use of dialect while the people of the south are more proud of speaking their own dialect. But we must also take into account, that we have the preconceived view in Germany that the middle and northern regions are considered as the regions where the standard German is spoken. This might have an influence upon the decision of a transcriber, as to whether a pronunciation should be marked as a deviation from standard German or not. The relatively moderate amount of deviations in the south Bavarian region "u" could be explained by the high number of speakers from the big city of Munich, who speak in a more urban colloquial speech. The opposite region "a" from the very north of Germany with a relatively high amount of deviation shows a known increase of dialect speaking people in the most northerly part of Germany, North-Friesland.

The results are similar to those of our experiment in [3], where our listeners had to recognize the speakers' origin by hearing the recordings of the RVG1 digit sequences. The recognition rates were high for Switzerland (83%), Austria (58%) and the southern part of Germany (37%). There might be a correlation between a high number of pronunciation variants and the perceptive recognition of a speaker's origin.

4.2. Topics and pronunciation variants

In the second part of the study we were concerned with the question, how far the topic of a speaker's monologue takes influence upon the frequency of pronunciation comments. We categorized our monologues with appropriate terms concerning the contents of the talk. We took very broad categories like "student life", "free time" or "job". Table 2 shows the results of the 6 topics 95% of the monologues are concerned with. 53,75% of our speakers were students, therefore, the most monologues were concerned with student life. These monologues lay upon the average of 23,55% variation. It seems, that if a speaker describes events of the job, he/she would use a more standardized speaking style (14% pronunciation comments). But it has to be taken in account, that we would need more monologues from a speaker with different topics for more reliable results.

category	frequency	variation
student life	20	25
free time	17	24
course of the day	16	20
job	14	14
mix	11	16
travel	7	21

Table 2: topic categories, frequency of monologues in percent, pronunciation comments in percent (variation)

4.3. POS and pronunciation variants

The final part of our study deals with the relation between linguistic word categories and pronunciation variants in the RVG1-monologues.

For this approach we summarized the POS tagged tokens into broad linguistic classes. For all the classes with a token frequency of more than 5% we counted the percentage of commented pronunciation per token. These results can be seen in figure 2.

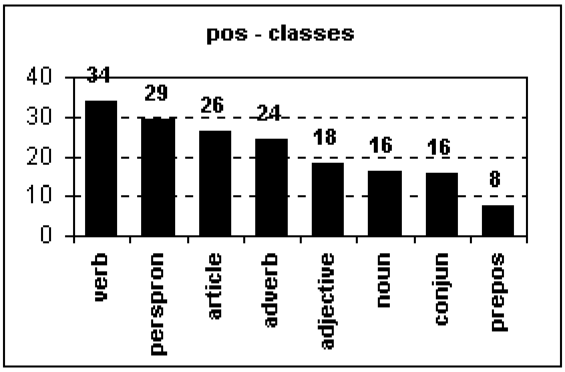

Figure 2: broad linguistic classes with a token frequency upon 5%; pronunciation comments in percent (see also table 1)

Further, we asked which of the single categories within the linguistic classes are more affected, which less. We found that some of the categories are very infrequently commented as a variant. For example, the proper nouns have only in 13% of appearance a pronunciation comment. Others are highly affected, e.g. the auxiliary past participle verbs on nearly 67%. The average of the percentage per POS category lays on 25%. Some of the word categories within the classes are not enough in use in our corpus for meaningful results, so that we had to leave them out.

Figure 3 shows the categories within the linguistic classes which appears more than one hundred times in the entire corpus and the percentage of how often they are commented on striking pronunciation. The black bars give the percentage of the frequency of the single POS category, the white bars show the percentage of commented pronunciation per POS category.

Accordingly to results from other studies (for example [1]), we found the highest pronunciation rates in the group of function words: auxiliary verbs (VAFIN 41%, VAPP 67%), pronouns (PPER 31%), articles (ART 26%) and prepositions (APZR 67%). But also the content verbs show considerably high results (VVP 38%, VVINF 27%). Less affected by pronunciation comments are nouns (NN 16%, NE 13%), and conjunctions (KON 17%, KOUS 12%), while adjectives (ADJA 20%) and adverbs (ADV 25%) can be found close to the average. As an additional result we found 78% pronunciation deviation at the negation particle for a frequency of 608 tokens within the entire corpus.

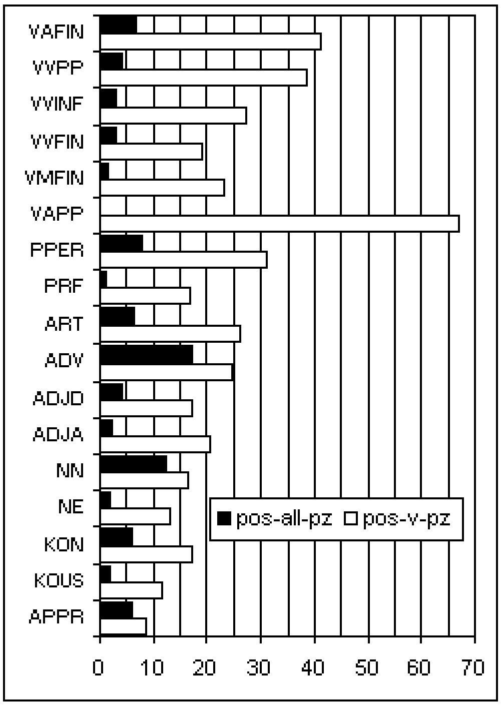

Figure 3: Pos categories, pos-all-pz = frequency of the category in percent (black bars), pos-v-pz = pronunciation comments per category in percent (white bars)

5. CONCLUSION

We showed that spontaneous speech within the German speaking regions differs regarding to the amount of pronunciation variations of the standard language and the region of origin. If the task of a database consists in the collection of regional variety (e.g. neither strong dialect nor standard language) one could take an average of 20% pronunciation variants in account. The RVG1 corpus shows an higher average due to the fact that also speakers of Switzerland and Austria had been recorded. Our results let assume that the German of these regions is far away from to be a part of standard German and could falsify the deviation rates. A further aspect of creating appropriate databases for analysis of a standard language should be the scenario of the recording. which means, what topic a speaker is ask for to speak about. The topic should be consistent for all the speakers. Even better would be if a speaker produces talks with different topics. Our results shows that the topic "student life" and the topic "job" produces speech styles of appropriately different level.

The results concerning the linguistic categories in relation to the pronunciation variants agree with former results that function words are highly affected.

As a further task, we will analyze the kind of variations we found in the material.

6. ACKNOWLEDGMENTS

We like to thank Martin Emele from the IMS, University of Stuttgart, who did the automatic POS tagging for our experiments. Further, this work couldn't have been done without the experiences and developments within the VERBMOBIL project.

7. REFERENCES

1 Beringer, N., Schiel, F., Regel-Brietzmann, P., "German Regional Variants – a Problem for Automatic Speech Recognition?", ICSLP '98, Sidney, 1998.

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

3 Burger, S., Draxler, Ch., "Identifying Dialects of German from Digit Strings", Proc. of LREC 1998, Granada, 1998.

4 Burger, S., Kachelrieß, E., "Aussprachevarianten in der Verbmobil-Transliteration - Regeln zur konsistenteren Verschriftung", Verbmobil Memo-111-96, München, 1996.

5 Burger, S., Schiel, F. "RVG 1 - A Database for Regional Variants of Contemporary German", Proc. of LREC 1998, Granada, 1998.

6 Duden - Rechtschreibung der deutschen Sprache, 20., neu bearb. und erw. Aufl. Dudenverlag, Mannheim, Wien, Zürich, 1991.

7 Feldweg, H., "Implementation and Evaluation of a German HMM for POS Disambiguation", From Texts to Tags: Issues in Multilingual Language Analysis. Proc. of the ACL SIGDAT Workshop, (pp. 41--46), Dublin, 1995.

8 König, W., Dtv-Atlas zur deutschen Sprache, Dtv-Verlag, München, 1978.

9 Schiel, F., "The Bavarian Archive for Speech Signals: Resources for the Speech Community", Eurospeech ’97, (pp. 1687--1690), Rhodos, 1997.