What are transcription errors and Why are they made?

Daniela Oppermann☆, Susanne Burger★, Karl Weilhammer☆

☆Institut für Phonetik und Sprachliche Kommunikation
Schellingstr. 3, 80799 München, German
{daniela.oppermann/karl.weilhammer}@phonetk.uni-muenchen.de
★Interactive Systems Laboratories, CMU Pittsburgh, USA
sburger@cs.cmu.edu

Abstract

In recent work we compared transcriptions of German spontaneous dialogues of the VERBMOBIL corpus to ascertain differences between transcribers and quality. A better understanding of where and what kind of inconsistencies occur will help us to improve the working environment for transcribers, to reduce the effort on correction passes, and will finally result in better transcription quality. The results show that transcribers have different levels of perception of spontaneous speech phenomena, mainly prosodic phenomena such as pauses in speech and lengthening. During the correction pass 80% of these labels had to be inserted. Additionally, the annotation of non-grammatical phrases and pronunciation comments seems to need a better explanation in the convention manual. Here the correcting transcribers had to change 20% of the annotations.

1. Introduction

Basically, a transliteration of spontaneous dialogues in VERBMOBIL (Oppermann &Burger, 1999) consists of:

- Orthographic word level transliteration, plus tags for several word classes (proper names, digits)
- Annotation of spontaneous phenomena by means of specially defined labels
- Annotation of background noises
- Structural information such as bracketing non-grammatical phrases

These transliterations have to serve different partners within the project as a basis for further annotations, training data, or simply as textual representation of the dialogues. A high consistency in the use of conventions (Burger, 1997; Burger & Kachelrieß, 1996)allows the partners to easily process the transcribed data. It makes results procured by different partners comparable. Since automatic transliterations of the same quality as manual transliterations are still not available, a certain amount of typical errors is always to be taken into account. On the other hand, though, even trained human transcribers tend to differ in their perception of the phenomena, or simply make mistakes in using the rules. Previous analyses revealed that despite well defined catalogues of transliteration rules and the quality of technical equipment, "the quality of speech annotations used for technical applications must be seen against the background of description level, inherent perceptual features of the speech sounds in a language, and the requirements of the performed labeling task" (B. Eisen, 1993).

Assistants with different educational background, mostly students of different faculties, and not necessarily students of a language science usually do transcription work. As long as transcription rules are intelligible and annotation tools easy to handle, the only skill a transcriber has to offer is appropriate orthographic knowledge of the language of the transcription and a good sense of hearing. As a precaution, all the VERBMOBIL transliterations went through a final correction pass (final pass) done by highly experienced transcribers before they were published. However, the comparison between the first pass version and the final pass version still results in a considerable amount of difference between the passes. To learn how we may reduce the correction effort by improving the first pass transcription, we want to know several things, such as, which kinds of inconsistencies occur within different states of transliterations. Where do they occur and why do they occur? In the present work we compared VERBMOBIL transliterations of these different levels (first pass and final corrected). Additionally, we analyzed a transliteration done by six different transcribers to find inter-individual differences within the data and unclear cases in the transliteration conventions.

Our hypothesis is that there are three different types of error sources:

1) Writing against familiar rules (i.e. unusual compound rules, the tagging of word categories such as digits and proper names)
2) Perception of events, which is secondary in normal speech perception (i.e. breathing, pauses, and special pronunciation)
3) Annotation rules, which are difficult to understand (i.e. a complex system for marking speaker-speaker interference or the annotation of non-grammatical phrases, which requires a deeper understanding of syntactical structures).

2. Data

Three different types of transliteration were chosen and compared.

Group 1:

50 first pass transliterations were compared with their final pass versions. Students mostly from other faculties made the first-passes. All have good hearing skills, are able to write correctly within orthographic rules, and worked more than half a year on this task. All used the same type of headphones and the same transcription tool. Two specialists who are training the transcribers and have done transcriptions for years made the final passes.

Group 2a:

Comparisons of first pass transcripts of one dialogue annotated by 6 different persons.

Group 2b:
Comparison of final pass transliterations of two dialogues, which had been accidentally corrected by two different persons.

A lot of errors can be checked and corrected automatically such as spelling errors or formal convention errors. We grouped the remaining inconsistencies, which have to be corrected by hand according to the error source categories we mentioned in the introduction.

3. Results

3.1. Comparison of first pass and final pass transliterations

We counted the average occurrence of phenomena every 1000 words. We compared the average amount of first pass and final pass transliterations in the 50 dialogues and grouped them together with the error source classes.

Generally, in all error categories the amount of annotated phenomena increased in the final pass. As can be seen in Figure 1, writing against common rules shows almost no remarkable difference between first and final passes.

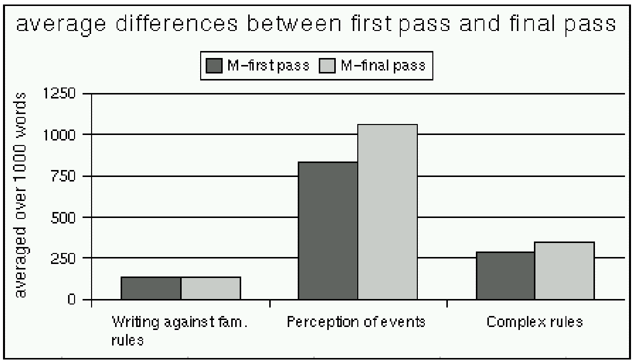

Figure 1: Average differences over all categories

A significant difference (20%) between the first and final pass can be seen for those labels where the transcriber had to perceive events which are secondary to speech understanding (pauses, abortions of articulation etc). The group of more complex annotation conventions also shows slight differences (8%).

3.1.1. Error analysis

In the next step we analyzed the differences in more detail to see what happened with the labeling of phenomena between the first and final pass, because even if there was no difference in the amount of errors it does not mean that nothing was corrected. As Tillmann & Pompino-Marshall (1993) already mentioned, four cases comparing two different stages of symbolic representation are found: identically, substitution, insertion and deletion.

Category 1: writing against familiar rules

In the first error category, "writing against familiar rules" errors occurred where the transcriber forgot to tag special word categories (proper names, digits, foreign words and neologisms) or had difficulties using a hyphen in longer compounds, which is not common according to German orthographic rules.

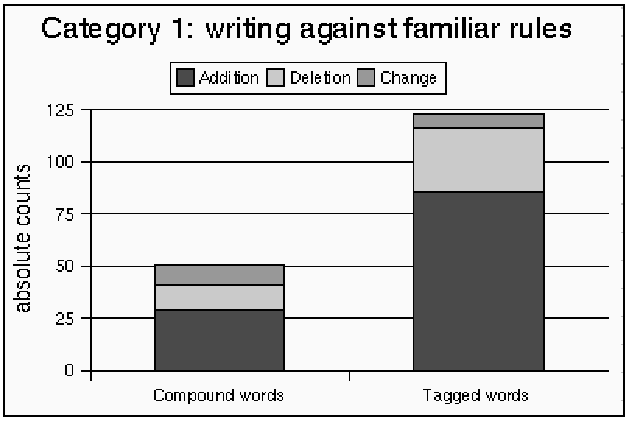

Figure 2: Category 1

4% of the differences between first and final passes occurred in compounds. More than half of these cases involved insertions (56%), and the rest were replacements (20%) or deletions (24%). In the cases where people had to tag special word categories only a 5% (123 cases in 2114 tags) difference could be found at the corrected versions. Most of the differences were insertions in the correction pass (70%), 24% had been deleted, and only a few (6%) were substituted by another tag.

Category 2: Perception errors

Compared to the other error categories, most errors occurred in the second category where the transcriber had to annotate additional spontaneous phenomena which are secondary in normal speech perception. To make these cases clearer we divided this group into three subcategories:

[a] Perception of phenomena occurring during articulation of words
[b] Perception of nonverbal speech phenomena
[c] Perception of noise

In the first case -- case [a] -- a transcriber has to tag the position of a word abortion, i.e. where a speaker doesn’t finish the articulation of a word and stops it at a special position, or the transcriber marks words or phrases which are not -- or mostly not -- identifiable. A listener is normally able to compensate these phenomena in normal speech. Therefore, a transcriber might overhear these cases.

What we found is that in general for more than half of the annotated phenomena corrections were required (52%). Figure 3 illustrates that, except for the not identifiable words, in every class more phenomena had to be inserted than deleted or changed.

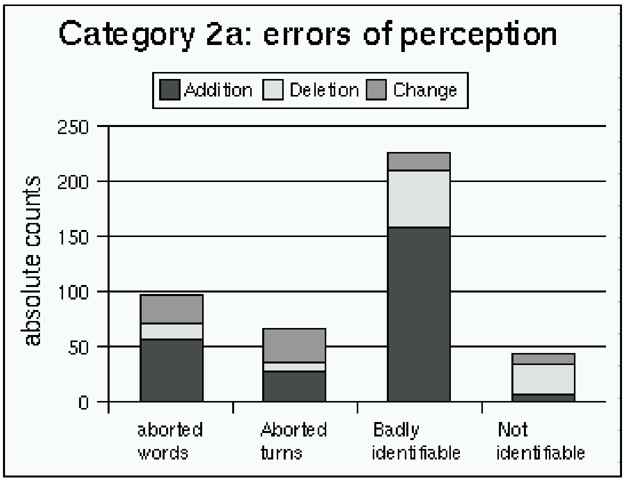

Figure 3: Category 2a

Case [b] contains spontaneous phenomena pertaining additionally to articulated speech such as speech pauses, breathing, filled pauses (hesitations) and lengthening of sounds within words.

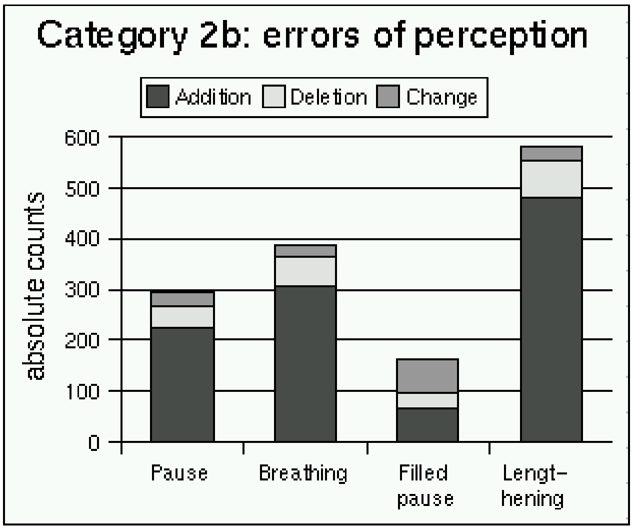

Figure 4: Category 2b

Counted over all phenomena found in the final pass texts, 24% had to be corrected. In general it can be said that the first pass transliterations still miss about 75% of the annotation of pauses, breathings and lengthening of sounds. About 20% had been wrongly placed and were therefore deleted, and about 5% of them were substituted. Hesitations seemed to be already placed quite correctly in the first pass, but were often replaced by another category of hesitation. For example, the category <”ah> (pure vowel) was corrected as <”ahm> (vowel plus nasal).

The third case [c] contains annotated noise phenomena. In VERBMOBIL we distinguish between two kinds of noise categories: human noise (laugh, cough, swallow, throat, smack and trash category) and technical noise (knock, rustle, squeak and trash category). All these phenomena were annotated when they are perceived between words and additionally at the same time of a word. 37% of the annotated noise phenomena were corrected in the final version.

Again it can be seen in figure 5 that most of the cases had to be inserted into the text (78%). Most errors occurred where noisy background was interfering with speech.

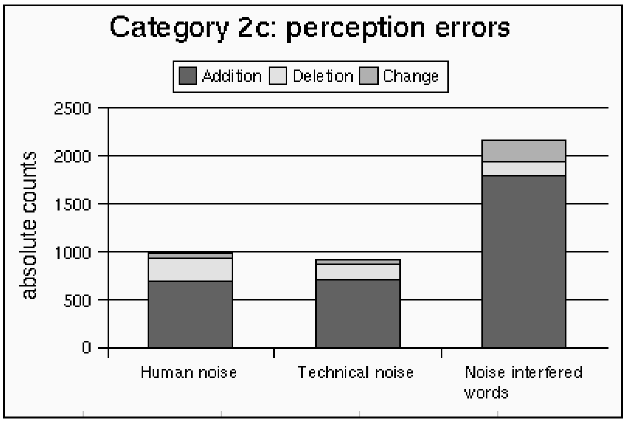

Figure 5: Category 2c

Category 3: complex annotation rules

Here we find rather complex convention rules, which are not easy to understand. This means that a transcriber has to understand the explanation and has to memorize more than just a label. There are three convention rules we consider as rather complex:

- Annotation of false-starts and repetitions
- Annotation of speaker-speaker-interference
- Annotation of transliteration comments, e.g. cases where the articulated words deviate from standard German.

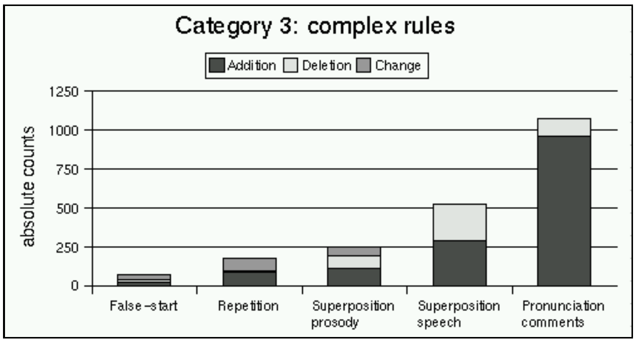

Figure 5: Category 3

Taken altogether more than half of the annotations were corrected (52%). False-starts were annotated when the speaker stopped in the middle of a sentence and started a new sentence without referring to what he said before. Repetitions were annotated when the speaker repeated or corrected phrases or words. As can be seen from Figure 5 the amount of inserted repetitions and false-starts are negligible in comparison to the other features of this category. A striking effect is that we have almost as many insertions as substitutions (about 45% each) in the annotation of this phenomenon and a few deletions. That lets us assume that transcribers do not have problems perceiving them, but that they are not sure what kind of label to use in the annotation of the perceived cases. This may indicate that the principle of this rule is not easy to understand. In the cases of speaker-interference we distinguished between superimposed prosodic phenomena (pauses, breathing, hesitations) and superimposed speech. Generally, we found many more cases of speaker-interfered speech than prosodic events in the transliterations.

In the last class of phenomena -- the annotation of transliteration comments -- we also found a relatively high amount of errors (34%), where most of them had to be inserted into the transliterations during the final pass (73%).

3.1.2 Summary

Generally, most annotated phenomena fall into category 2 - perception of events. In all categories more phenomena had to be added than deleted or replaced.

Most phenomena of category 1 - writing against familiar rules - were inserted during the correcting of the transliterations (63%).

In the case of perceptual phenomena most errors occurred in noise annotation. Generally more than half of the events were inserted. Some of the categories show some exceptions: In the case of not identifiable words the correcting person deleted 63%. Aborted articulations have as many substitutions as insertions (45%).

The annotation of filled pauses differs considerably from the others in its class. While the overall pattern of the category of prosodic events shows an average of about 80% insertions, hesitations taken alone the same amount of labels have been replaced as well as added into the transliterations.

In the category of noise annotation again most cases were inserted. During pauses between words 60% noises were added and in case of word interfering noises about 80% noises were inserted.

Altogether, we can assume, that most errors occur due to perceptual factors in the first pass transliterations.

3.2. Inter-transcriber differences of first pass transliterations

A transliteration done by six different transcribers on first pass level gave an impression about inconsistencies between transcribers at the same level.

Again, we split the counted phenomena into the source error. We concentrated on those phenomena that were annotated differently by at least four of the six transcribers.

We will only display absolute numbers in the following diagrams because the number of occurred phenomena in this single dialogue was so small.

In the error source categories 1 and 2[a], where the transcribers had to tag on word level, all transliterations have been transliterated almost consistently.

Differences could be found in those cases where the transcribers had to pay attention to phenomena other then speech, such as noise or prosodic phenomena and additionally, for non-grammatical phrases and deviations from standard German. The following will show a more detailed view of each of these categories.

Figure 6 shows that the first pass transcribers differ in category 2[b] only in the annotation of pauses and breathing. Not one of the transliterations of the same dialogue had the same number of annotated breathings or pauses in common with another. Besides transcriber G043ac_rob, the annotation of lengthened sounds appeared quite consistently.

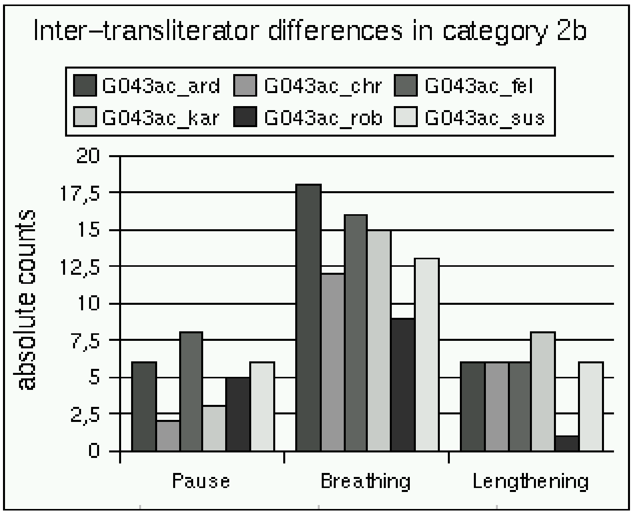

Figure 6: Inter-transcriber differences of first pass

Category 2[c] -- the noise annotations -- in figure 7 shows remarkably more inconsistencies between the transcribers. There seems to exist similarities in the pattern of distribution in the categories technical noise and noise interfered words.

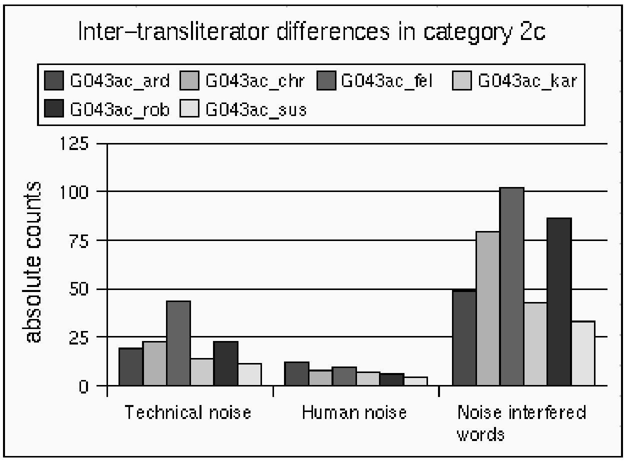

Figure 7: Inter- transcriber differences of first pass

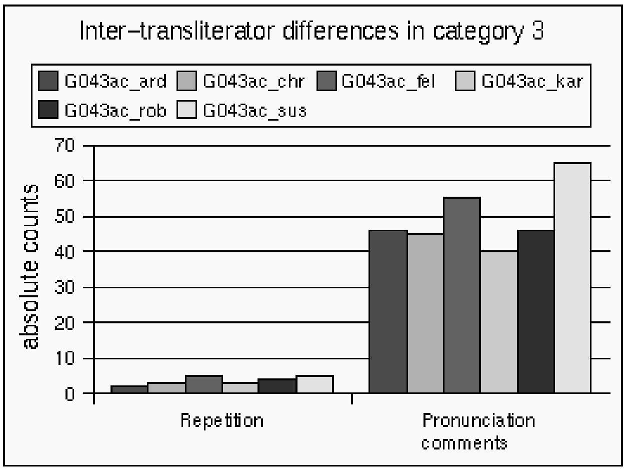

Figure 8: Inter- transcriber differences of first pass

Figure 8 displays the inconsistencies of annotations of category 3, repetitions and pronunciation comments. All of the transcribers agreed in the annotation of false-starts occurring in this dialogue; therefore, false-starts are not displayed in the diagram. It is remarkable that there are differences in the number of annotated repetitions/ corrections, given the agreement in annotation of false-starts. Inconsistencies in the number of annotated pronunciation comments show again that the decision on commenting or not depends more on individual opinion than if a general rule would exist.

Summary

Strong inconsistencies among the transcribers can be seen in the following cases:

- Pauses and breathing
- annotation of all noise categories
- repetitions of words and phrases and pronunciation comments

In these categories every transcriber labeled a different amount of phenomena.

3.3. Inter-transcriber differences of final pass transliterations

In our last analysis, we compared the corrected transliterations of two dialogues that accidentally went trough the final pass of two different correctors.

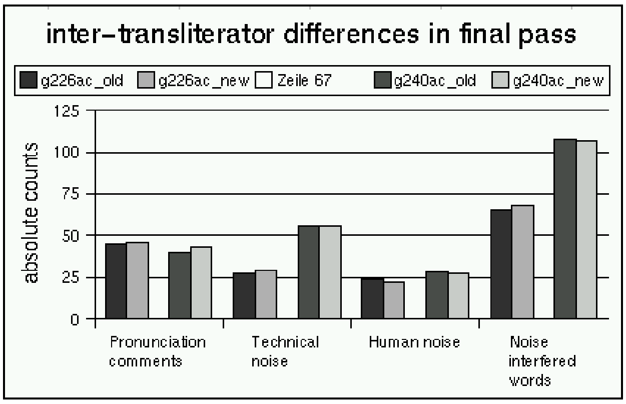

Figure 9: inter- transcriber differences of final pass

First of all, figure 9 shows only very small inconsistencies between the corrected versions. However, differences still can be found for noise annotation and pronunciation comments.

Summary

The differences are very small, which shows that consensus in annotation is possible, but there are cases where even experienced transcribers with the same kind of training disagree.

4. General Discussion

The high number of insertions of annotations in contrast to changes or deletions shows that a lot of phenomena escaped the transcribers’ notice. One reason might be an individual threshold for the decision if a perception is worth an annotation or not. On the other hand there are phenomena which are difficult to perceive. It might depend on the training and experience of a transcriber if a phenomenon is perceived at all.

There are different levels of perception tasks a transcriber has to fulfill. S/he has to transcribe the spoken words, add special tags, and listen to the background. Often this cannot be done stepwise, since time and money plays an additional role.

The small number of deletions and substitutions indicate that most phenomena were perceived and placed correctly in the first pass.

Besides the large number of insertions, there are additional inconsistencies, which may represent difficulties transcribers have with certain transcription rules. Some interesting cases we found are discussed in the following:

Compounds: An explanation for the large number of inconsistencies in hyphening compounds may be that the German language allows longer word compounds without using a hyphen. Inserting hyphens requires an identification of the word parts, which in normal writing is not considered. Here the transcribers clearly have to handle unfamiliar rules. The tagging of special categories -- also unusual -- seems to be easier. This is, however, not the change of an old rule but a completely new task.

Non-identifiable utterances: We often found cases where an annotated non-identifiable element is identified later by the corrector. Also here there might have been perception problems due to background noises within the transcription lab. Or the part a transcriber listened to was too restrictively selected, so that context information could not help. Experienced transcribers might have better concepts for the identification of utterances which are difficult to identify because of bad articulation or dialect.

Hesitations/filled pauses and noise: The category of filled pauses shows as many replacements as insertions. Besides not perceived hesitation, also the annotated category of a hesitation often had to be corrected. Here a perception problem might be the reason. But there is also the possibility that transcribers have difficulties sorting hesitations into the correct categories of hesitation offered in the manual (such as vowel-like, vowel plus nasal, nasal and trash category). In the case of noise, this effect may be even stronger; noise might be processed differently in perception. Here too, a transcriber has to sort a perceived noise into a special noise category and that might be difficult in a lot of cases.

Non-grammatical: The category false-start and repetition had a relative high rate of substitutions, which may indicate problems in the understanding of the correct usage.

The small number of deletions and replacements, but high number of insertions in the category of pronunciation comments indicates that the transcribers had no problems of how to make a comment, but they might have had problems deciding when a comment was necessary. This effect might be due to the transcribers' regional origin and the threshold when they thought a spoken word deviated far enough from the standard language to be worth a comment.

The inter-transcriber comparison leads to similar results:

Differences in pauses, breathing and also in noise annotation may arise from individual thresholds in the perception. The same individual concepts may also play a role in the differences in the annotation of pronunciation comments and repetitions. Additionally certain convention rules seem not to be clear enough to result in consistent transliterations.

5. Conclusion

The goal of this study was to improve the working environment of transcription work and to reduce the effort spent on correction passes. We hoped to find the transcription errors by analyzing the differences between first pass transliteration and final pass transliteration.

We found inconsistencies in the annotation of all phenomena in all transcriber groups. The following three points might explain why these differences occur in the transliteration of spontaneous speech.

First, there is a large variety of different phenomena to be annotated. In the contrary to this number, the less time and money spent on transcriptions requires a certain speed, which might result in not-perceived phenomena and omitted annotations.

Second, some convention rules might not be explained well enough in the transcription manual. The results show that especially the definition for repetition/false starts should be updated. A better explanation of noise and hesitation categories might be helpful.

Third, there still remain some phenomena which probably could never be consistently annotated by human transcribers due to the fact that they are based on individual perception. If these annotations serve as training data for i.e. a breathing model where it would not be necessary that all occurrences of them be really annotated, then our data shows that most of them are at least annotated always the same way. There were not so many replacements or deletions found in the corrected transliterations. If a consistent annotation of all occurring phenomena is required then the question remains if those annotations make sense at all.

6. References

Burger, S. (1997). Transliteration spontansprachlicher Daten - Lexikon der Transliterationskonventionen - VERBMOBIL II. *Verbmobil Tech-Dok-56-97*. München.Germany.

Burger, S., Kachelrieß, E. (1996). Aussprachevarianten in der Verbmobil Transliteration - Regeln zur konsistenteren Verschriftung. *Verbmobil Memo-111-96*. München. Germany

Oppermann, D.,Burger, D. (1999). What Makes Speech Data Spontaneous? *Proceedings of the ICPhS 1999.* San Francisco. USA.

Eisen, B. (1993). Reliability of Speech Segmentation and Labelling at Different Levels of Transcription. *Proceedings of EUROSPEECH 1993 (pp. 673 - 676).* Berlin. Germany.

Tillmann, H.G., Pompino-Marschall, B. (1993). Theoretical Principles Concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems. *Proceedings of EUROSPEECH 1993 (pp. 1691 - 1694).* Berlin. Germany.

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