Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus

Simo Goddijn† **and Diana Binnenpoorte**‡
† Speech Processing Expertise Centre, University of Nijmegen, The Netherlands
‡ A^{2}RT, Department of Language and Speech, University of Nijmegen, The Netherlands
E-mail: S.Goddijn@let.kun.nl, D.Binnenpoorte@let.kun.nl

ABSTRACT

For research and development purposes in the areas of phonetics and speech technology, phonetically transcribed speech may be of great value. In the near future, the Spoken Dutch Corpus (CGN) is going to offer such transcriptions for about one thousand hours of spoken Dutch, of which 90% will consist of automatic transcriptions and 10% of manually produced transcriptions. An advantage of automatically produced transcriptions is that they are maximally reliable; they are however not necessarily maximally accurate. One way of making them more accurate is having them checked and modified manually, but it is widely accepted that human transcriptions tend to be subjective and unreliable. The goal of this paper is to establish if human CGN transcribers succeeded in making accurate transcriptions by correcting automatic transcriptions, while maintaining a high level of reliability.

1. INTRODUCTION

The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) is a Flemish-Dutch initiative aimed at the compilation of a large (10 million word) corpus of spoken Dutch that will contain speech from a great variety of socio-situational settings. This project will create an important resource for research in various linguistic disciplines and for developments and applications in language and speech technology (for further details, cf [1].) All speech material in the corpus will be orthographically transcribed, lemmatized and enriched with part-of-speech information. For about 1 million words more detailed information will be provided, such as a syntactic annotation, a hand-checked word segmentation and a broad phonetic transcription.

In this paper a thorough explanation is given of the choices that were made with respect to the level of detail in the broad phonetic CGN transcriptions and the transcription procedure.

Transcriptions by human listeners are known to be subjective and not very reliable [2]. Automatically generated transcriptions on the other hand, are very objective and reliable. They are usually obtained by translating orthographic symbols into phonetic symbols using a lexicon with phonetic transcriptions for every word. Several kinds of phonological rules may be applied to the outcome to account for phonological processes that occur in connected speech. For many phonetic research purposes however, this type of automatic transcriptions is not ideal, since they only reflect the expected pronunciation variation, not the pronunciation variation that has actually been realized. In terms of cost and consistency, the best option would be to generate an automatic transcription by taking into account the acoustic signal. It is, however, difficult to account for all of the pronunciation variation that is characteristic of especially more spontaneous speech varieties. One possibility would be to create a lexicon with pronunciation variants for every word and to perform so-called forced recognition: an automatic speech recognizer with knowledge of the orthographic transcription is forced to choose between the phonetic variants of every orthographic word and then picks the variant that most closely matches the sound file [3]. The knowledge needed for the creation of such a lexicon, however, is precisely the kind of knowledge we lack. Much knowledge is available about phonological processes, but little is known about how these processes behave in true spontaneous speech. Therefore, it was decided that the broad phonetic CGN transcriptions should be made by hand. One of the most direct purposes of the hand-made transcriptions will be to ameliorate the automatic transcription procedure in order to provide the nine million CGN-words that are not manually transcribed with a good phonetic transcription.

For phoneticians and speech technologists, in order to be able to make sensible use of the data, it is essential to know what can and what cannot be expected of the transcription data, given the transcription procedure adopted in the CGN, and explained in detail in section 2. In short, multiple transcribers process data on the basis of automatically generated provisional transcriptions. In this paper we report on research aimed at assessing the quality of the obtained transcriptions. In [4] we reported on the accuracy of the transcriptions by comparing them to a reference transcription (see 3.3). In this paper we go on to make an analysis of the “errors” in relation to the reference transcription and we give insight in consistency between different transcribers. Furthermore, by establishing the number of changes made by the human transcribers to the automatic transcriptions, we try to demonstrate the validity of the adopted procedure. By means of this paper we intend to contribute to the usability of the phonetic CGN transcriptions.

ISBN 1-876346-48-5 © 2003 UAB

2. THE TRANSCRIPTION PROCESS

2.1 The level of detail

Phonetic transcriptions can be made on various levels of detail. The choice for broad phonetic transcriptions was made for a number of reasons. First, the higher the level of required detail, the bigger the risk of disagreement between transcribers. For example, in [5] a mutual agreement of only 33% is reported for the use of diacritic symbols. Second, the higher the level of required detail, the more time (and money) the transcription work will cost. From a pilot study on broad transcriptions, it appeared that transcription time of one minute of speech varied from 35 to 60 minutes, depending on the transcribed speech variety. Probably (much) more time would have to be spent to obtain narrow transcriptions. Third, it is a problem to find qualified transcribers. It was felt that broad transcriptions could possibly be made by students, whereas narrow transcriptions would require the experience of an expert phonetician. Apart from the increase in transcription cost that would imply, it would be very hard to find phoneticians who are willing to spend a great deal of time to this routine job. Last, it is not easy to decide on the kind of detail that should appear in a narrow transcription. Details that are important to one linguist may hamper the research of another. The combination of the above mentioned factors would not justify a choice for narrow transcriptions.

The transcription symbols that are used in the CGN project are based on SAMPA [6]. The set contains 46 symbols among which all voiced/voiceless contrasts of Dutch plosives and fricatives are represented. In the symbol set a distinction is made between /x/ and /G/, the first symbol representing the voiceless uvular fricative and the second symbol its voiced variant. The distinction is justified by the fact that in some southern parts of the Netherlands, as well as in Flanders, a distinction between the sounds is made and experienced in spoken language. In the rest of the Netherlands however, both speech sounds are used, dependent on idiosyncracy and context, but there is no awareness of the distinction. It proved to be very difficult to make our transcribers aware of the distinction, especially since they tend to confuse it with another difference in pronunciation of /x/ between the southern and northern part of the Netherlands, which is much more conspicuous to them (i.e. the more velar pronunciation of this speech sound in the south).

2.2 Quality assurance

Bearing in mind that human transcriptions are susceptible to unreliability and subjectivity, a number of precautions have been taken to try and minimize these risks. For example, transcribers are supervised by a phonetician who monitors the transcription process closely, especially during the training period. Recurring mistakes are detected and discussed and an attempt is made to agree on phoneme categories. For Dutch it appeared that the voiced-voiceless distinction requires special attention. Students are required to work in the same room to be able to consult with each other. They are only hired if they are willing to participate in the project for at least 12 hours a week for a period of at least half a year. This procedure is followed because it is believed that the fewer students work on the project, the better it is for consistency's sake. Furthermore, for half of the transcribed data (with a priority of spontaneous speech over other speech components) a second transcriber corrects the work of the first one.

2.3 Automatic transcriptions as starting point

A pilot study showed that human transcriptions were most efficiently made if transcribers do not start from scratch, but modify an automatically generated transcription (the AT) until it reflects what actually has been said. An additional advantage of this procedure is that it provides a solution for cases of doubt: whenever there is doubt between two symbols, transcribers are required to leave the symbol from the example transcription, thus improving reliability. In adopting this procedure, there is a certain risk of creating a bias towards the AT. Therefore, it is expressly pointed out to transcribers that they should consider the AT as no more than it is - a means to save typing time and to help prevent typing mistakes. Transcribers are encouraged to change anything that does not correspond with the speech signal.

3. METHOD

3.1 Speech material

The speech material used in the experiment consists of 16 minutes of speech, containing 2712 words. This subcorpus extracted from the CGN contains five one-minute samples of read speech (RS) and lectures (LC) and three one-minute samples of interviews (IN) and spontaneous conversations (SC). The samples were chosen so as to vary with respect to speakers' sex, age and region of education. Thus a representative sample of Northern Dutch was obtained.

3.2 Manually corrected transcriptions (HT)

The manual transcriptions for this experiment (human transcriptions, HTs) were produced in exactly the same way as the 'real' CGN transcriptions. Four transcribers (language students) who had all been working on CGN transcriptions for more than five months each transcribed the complete 16 minutes. Transcriptions were made using the interactive speech processing tool PRAAT [7], that allows users to listen to the speech signal and enter the transcription simultaneously. Although transcribers had an oscillogram of the speech signal at their disposal, they were instructed to make auditive transcriptions and not to revert to visual information.

An automatically generated transcription (AT), in which all so-called obligatory word internal processes are applied (for an elaborate description, see [8]), was corrected by hand according to the rules of a written protocol. In the protocol is stated that phonetic processes (insertions, deletions and substitutions) that would result in a sound represented by a different SAMPA symbol must be reflected in the transcription. Gradual processes like degree of voicedness in plosives and fricatives or monophthongising in vowels are not expressed, because the symbol set does not contain diacritics. The spontaneous speech samples (SC) were corrected a second time by a different transcriber, just as in CGN practice.

3.3 Reference transcription (RT)

As a reference, a consensus transcription was used that was established by having two expert listeners agree over every transcribed symbol. No AT was made available to the experts, but they did have the orthographic transcription at their disposal.

3.4 Alignment of transcriptions

All transcriptions revised by the four transcribers were to be compared with the reference transcription. This was performed with the program Align [2], which uses a dynamic programming algorithm to make an alignment between two transcriptions on phoneme level. A distance measure is calculated by Align on the basis of articulatory features like place and manner of articulation, voice, lip rounding, length, etc. For example, the distance between /t/ and /d/ is smaller than the distance between /t/ and /x/. The distances are used to calculate the optimal alignment. Deletions and insertions always generate the same distance.

In previous research [4], Align was used to compare each of the four HTs to the RT to measure transcription accuracy. A detailed analysis of the results is given in this paper. Furthermore, an alignment was set up to establish the number of changes the transcribers make to the ATs to obtain their HTs. Finally, a series of alignments were conducted to find out to what extent the four HTs agree with each other.

4. RESULTS

4.1 Accuracy

From previous experiments [4] it became clear that HTs differ more from the RTs for spontaneous speech styles than for prepared speech styles. This tendency was shown for every transcriber. The percentages agreement ranged from 93.9% for RS to 85.3% for SC. There was no large variation between transcribers in the degree of deviation from the RTs except in the SC condition. That is explained by the fact that the SC samples were corrected by a second transcriber. The correction cycle appeared to have a positive effect on transcription accuracy.

An analysis of the errors shows that the majority originated from substitutions. For the more spontaneous speech styles, substitutions constituted around 50% of the errors and for RS around 70%. To achieve a better understanding of the nature of the deviations, a qualitative analysis was conducted. It appeared that on average 43% of the substitutions were caused by voice confusions. The remaining substitutions concerned for example vowel confusions and nasal confusions. In addition, several occasional substitutions were found for especially the more spontaneous speech styles, eg. /n/ for /d/. Substitutions of an unvoiced plosive or fricative into its voiced equivalent were more frequent than the other way round.

It was decided to focus on the five most frequently occurring confusions, which all concerned the substitution of an unvoiced plosive or fricative into its voiced equivalent. In Table 1 details are given about the proportion of all occurrences of the unvoiced phonemes in the RT that are substituted with their voiced counterparts in the HTs. The figures show that even for the phonemes that are most liable to confusion, /x/ in IN and /k/ in SC, 78% of the occurrences are transcribed in accordance with the RT.

%	x,G	t,d	f,v	s,z	k,g
RS	19	6	11	5	11
LC	13	10	5	8	11
IN	22	11	8	10	17
SC	20	10	21	13	22

Table 1 Percentage unvoiced-voiced substitutions of all occurrences of the unvoiced phoneme in RT

A closer look was taken at the different contexts in which the confusions appear. At least 84% up to 95% of the five most frequent substitutions took place in a context in which both the previous and the subsequent phoneme are voiced.

4.2 Agreement between transcribers

In order to rule out the possibility of high agreement percentages due to transcribers changing almost nothing to the AT, percentages of changed symbols were calculated. For RS the average percentage is about 10.5%, for LC 14.5%, for IN 17.1% and for SC 22.5%. In our opinion, these figures justify the transcription procedure in which transcribers correct an AT.

Table 2 gives an overview of percentage agreement between transcribers. Table 3 shows the distances calculated by Align, corrected for the number of phonemes to make them comparable between speech styles.

%	RS	LC	IN	SC
HT2/HT1	93.8	89.2	87.7	85.8
HT3/HT1	95.2	90.6	88.9	90.2
HT4/HT1	94.9	89.7	88.0	85.7
HT3/HT2	95.6	91.8	90.9	88.3
HT4/HT2	96.1	91.6	91.3	94.9
HT4/HT3	96.3	92.1	91.7	87.9

Table 2 Percentage of intra-transcriber agreement

The same trend already shown in [4] becomes clear from these tables: the more spontaneous the speech styles, the less agreement between transcribers. The relatively high agreement percentages for HT3/HT1 and HT4/HT2 in SC are again due to the fact that HT3 corrected HT1 and HT2 corrected HT4 in this condition.

The distance measures in Table 3 show that when there is more agreement, the distance is usually smaller. There is however no complete one-to-one correspondence: compare the agreement percentages of 94.9% of HT4/HT1 in RS and

 ISBN 1-876346-48-5 © 2003 UAB

HT4/HT2 in SC with distances of 8.7 and 12.5 respectively. This means that the differences between HT3/HT1 had less acoustic features involved than the differences between HT4/HT2. In the RS condition almost every substitution was due to difference in voicing, which is only one feature, while in the SC condition substitutions were more often due to more than one feature (e.g. /A/ and /@/, differing not only on the high-low dimension but also on the front-back dimension).

	RS	LC	IN	SC
HT2/HT1	11.2	24.4	30.2	32.6
HT3/HT1	8.7	20.9	27.9	23.1
HT4/HT1	8.7	23.3	30.6	32.7
HT3/HT2	8.3	18.6	22.1	26.5
HT4/HT2	6.6	18.7	19.9	12.5
HT4/HT3	6.8	18.4	20.5	27.8

Table 3 Distance corrected for # phonemes

4. SUMMARY AND CONCLUSION

One of the goals of the current investigation was to give insight in the accuracy of the manually corrected phonetic CGN transcriptions. To be able to make statements about accuracy, a reference is needed to which the transcriptions under consideration can be compared. Although it is clear that there is no such thing as a perfect transcription, the assumption was made that the reference transcription represents “the truth”. Comparing the HTs to the RT, it was found that agreement decreases with the spontaneity of the speech style and ranges from 94% in RS to about 85% in SC. The largest share of the inconsistencies was caused by substitutions. On average, 43% of these substitutions appeared to be caused by voice confusions. Substitutions of voiceless plosives and fricatives into their voiced equivalents were more frequent than substitutions the other way round. This is partly explained by the fact that voiceless plosives and fricatives are more frequent than voiced ones. However, the differences are larger than expected on the basis of this alone, especially for the more spontaneous speech varieties. From personal communication with transcribers it appeared that they were inclined to transcribe the voiced variant whenever a plosive or fricative was unclear or soft (although the instruction was only to pay attention to the feature voice, and not to other differences between voiced and voiceless sounds). For the voiceless phonemes that were most susceptible to substitution into their voiced counterparts, the agreement with the RTs was still 78% to 95%. For 84% up to 95% of these substitutions, the previous and subsequent phonemes were voiced. Apparently, transcribers found it more difficult to establish voicelessness in an all-voiced context. The phonemes that proved to be most susceptible to confusion, /x/ and /G/, are not distinguished in most northern Dutch speech varieties. Three of our four transcribers had trouble in discriminating these speech sounds: the /x/-/G/-confusions must be attributed entirely to these three.

A comparison between the AT and HTs showed that transcribers change 10.5% to 22.5% of the symbols in the AT, thus no reason was found to doubt the validity of the correction procedure. The second goal of this paper was to give insight in the agreement between transcribers. It is not possible to project transcription agreement directly to quality as long as the nature of the disagreements is not known, so we also used distance measures that take acoustic features into account. Percentages agreement were found to range from about 96% to about 86%, decreasing with speech spontaneity. Distance measures developed in a similar way, although they did not show a one-to-one correspondance with agreement: no relevant systematic differences were found between transcribers. This implies that the differences in the gravity of the substitutions made by the transcribers are not extremely large.

ACKNOWLEDGEMENTS

This publication was supported by the project "Spoken Dutch Corpus (CGN)", which is funded by the Netherlands Organization for Scientific Research (NWO) and the Flemish Government.

REFERENCES

[1] Oostdijk, N., "The Spoken Dutch Corpus: Overview and first Evaluation", *Proceedings LREC*, Athens, 887-893, 2000.

[2] Cucchiarini, C. *Phonetic transcription: a methodological and empirical study*, Ph.D. thesis, University of Nijmegen, 1993.

[3] Kessens, J., Wester, M. and Strik, W., “Improving the erformance of a Dutch CSR by modeling within-word and cross-word pronunciation variation”, *Speech Communication 29*, 193-207, 1999.

[4] Binnenpoorte, D., Goddijn, S. and Cucchiarini, C., “How to improve Human and Machine Transcriptions of Spontaneous Speech”, ISCA+IEEE workshop on spontaneous speech processing and recognition, Tokyo, 2003 (in press).

[5] Shriberg, L.D. en L. Lof. “Reliability studies in broad and narrow phonetic transcription”, *Clinical Linguistics and Phonetics*, 5, 225-279, 1991.

[6] SAMPA, developed by Wells, J. http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[7] PRAAT, developed by Boersma, P. and Weenink, D., downloadable from http://www.fon.hum.uva.nl/praat.

[8] Cucchiarini, C., Binnenpoorte, D. and Goddijn, S., “Phonetic Transcriptions in the Spoken Dutch Corpus: how to combine Efficiency and Good Transcription Quality”, *Proceedings Eurospeech*, 1679-1682, 2001.