Phonetic Transcriptions in the Spoken Dutch Corpus: how to Combine Efficiency and Good Transcription Quality

Catia Cucchiarini, Diana Binnenpoorte and Simo Goddijn

A^{2}RT, Department of Language and Speech, University of Nijmegen, The Netherlands

{c.cucchiarini, d.binnenpoorte, s.goddijn}@let.kun.nl

Abstract

This paper reports on an experiment aimed at establishing how phonetic transcriptions for the large CGN corpus can be obtained most efficiently. This experiment explores the potential of an automatically generated transcription (AGT) by comparing an AGT with a reference transcription (Tref) of the same material, to determine whether and how the AGT can be improved to make it more similar to Tref. The results indicate that the AGT can be optimized through pronunciation variation modelling so as to make human corrections more efficient or even superfluous, at least for some speech styles.

1. Introduction

The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) project is a joint Flemish-Dutch initiative aimed at the compilation of a large (10 million words) corpus of spoken Dutch that will contain speech from a great variety of socio-situational settings. The rationale behind this project is to 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 we focus on the phonetic transcription and on the decisions that must be made in this respect.

Phonetic transcriptions of speech, which are obtained by analyzing utterances auditorily into sequences of units represented by phonetic symbols, are extremely time-consuming and therefore costly. Moreover, phonetic transcriptions tend to contain an element of subjectivity and thus may differ when they are made by different transcribers (see [2] and references therein). With a view to improving transcription reliability, researchers often decide to have more than one person transcribe the same speech material, with obvious consequences in terms of time and costs. When the speech material to be transcribed is such a huge amount as the 1 million words of the CGN, making phonetic transcriptions is almost prohibitive. In cases such as this one, it is important to choose a procedure that ensures that an acceptable level of transcription quality is achieved at reasonable costs.

For all the reasons mentioned above, the CGN steering committee decided that an experiment be carried out to determine how phonetic transcriptions could best be produced in the project. Instead of designing an experiment in which human transcribers are asked to make phonetic transcriptions under various conditions, we decided first to explore the possibility of having the phonetic transcriptions produced automatically by a computer. In view of the recent developments in language and speech technology, this idea is worth investigating, before embarking on costly enterprises with human transcribers, which do not necessarily guarantee high quality transcriptions. If automatically generated phonetic transcriptions appear to be of satisfactory quality, they are a serious alternative to human transcriptions, with substantial reductions in time and costs. Automatically generated transcription also guarantee maximal consistency, which is also important when creating such large corpora as the CGN. The rest of this paper will be devoted to presenting this experiment and the results that have been obtained so far.

2. The phonetic transcription experiment

In the framework of the CGN project a list is compiled of all words contained in the corpus, the CGN lexicon, which provides the orthographic and the corresponding phonetic representation (obtained with TREETALK [9]) of each word. As for each utterance in the corpus an orthographic transcription is available, a corresponding phonetic transcription can be obtained by a simple lexicon-lookup procedure. At some points this automatically generated transcription (AGT) is likely to deviate from what was actually realised, and the deviations will probably be more substantial for spontaneous speech than for read speech.

Within the context of the CGN, such an AGT could be used in several ways, depending on the degree of deviation between the AGT and what was actually realised. For those cases in which the deviation between the AGT and the actual realisation is substantial, one could use the AGT as a basis transcription to be manually improved by human transcribers. The better the quality of the AGT, the more efficient the revision process will be. This procedure should result in considerable gains in efficiency, provided that the AGT is of reasonable quality, because if transcribers have to correct almost all symbols, it probably takes less time to transcribe from scratch. For the cases in which the deviations between the AGT and actual realisation are marginal, one could consider using the AGT ‘as is’ without human intervention.

Of course one could also consider the possibility of optimizing the AGT obtained by lexicon-lookup until it is of sufficient quality for as many speech types as possible. Instead of using canonical representations of isolated words, one could take an AGT in which, for instance, pronunciation variation is modelled, thus obtaining an AGT that comes closer to what was actually realised.

The experiment we decided to carry out was aimed at determining how much and in what respects an AGT obtained by lexicon-lookup differs from what was actually realized, for various types of speech. The information thus obtained can then be used to optimize the AGT in various ways, until it has reached a sufficient quality level so as to make human intervention more efficient or even superfluous.

In an experiment of this kind there are several important concepts that have to be defined and measured, such as, for instance, "what was actually realised", transcription quality, satisfactory transcription quality etc. Since the aim of the research reported in this paper is to find out how much and in what respects an AGT differs from what was actually realized, it follows that the first two concepts are particularly important here. Therefore in 3.1 we indicate how they have been operationalised in this experiment. The definition of satisfactory transcription quality will not be addressed here, because this is clearly beyond the scope of the present paper.

3. Method

3.1. Operationalisation of important concepts

The first thing to be defined in this experiment is how we are going to determine "what was actually realised", i.e. how we arrive at a phonetic transcription of the speech material that can be used as reference to evaluate the AGT. In phonetic research the difficulties in obtaining such a transcription are well known, and it is generally acknowledged that there is no absolute truth of the matter as to what phones a speaker produced in an utterance [3]. Hence there is no reference transcription that can be considered correct. To try and circumvent this problem as much as possible, phoneticians have been looking for procedures that can approach a reference transcription, such as a transcription made by two or more transcribers after they have agreed on each individual symbol: a consensus transcription [4]. This is the procedure that has been adopted in the present experiment to obtain a reference transcription (Tref) that can be used to evaluate the AGT.

Another thing that has to be defined is how to measure transcription quality. Given that we have a reference transcription, the obvious choice would be to carry out some sort of alignment between the reference transcription and the AGT, with a view to determining a distance measure which will also provide a measure of transcription quality. We will use a dynamic programming algorithm in which the distance between corresponding phonetic symbols is calculated on the basis of articulatory features defining the speech sounds the symbols stand for [3]. In addition to aligning two transcriptions, this algorithm compares the two transcriptions and returns various data such as an overall distance measure, the number of insertions, deletions and substitutions of phonemes, and more detailed data indicating to which features substitutions are related. In the present experiment this kind of information is extremely valuable to establish how the AGT differs from Tref, and, in turn, to determine how the AGT could be improved to make it more similar to Tref.

3.2. Speech material

The present experiment was limited to the Dutch language varieties spoken in the Netherlands. The speech material selected varies with respect to speech style and speaker (see 3.3), thus constituting a representative sample of the Northern Dutch part of the CGN, and consists of 16 different fragments representing four categories: read speech (RS), lectures (LC), interviews (IN), and spontaneous conversations (SC). A total of about 16 minutes of speech, containing 2712 words, was transcribed.). Table 1 provides details for the four categories.

Table 1 Overview of the speech material

speech style	# of speakers	# of words	duration
RS	1	682	04:57 min
LC	1	892	05:09 min
IN	2	523	03:01 min
SC	2	615	03:01 min

3.3. Speakers

The speech material was produced by twenty different subjects, nine female and eleven male speakers, who varied with respect to region of origin and age. Thus a representative sample of Northern Dutch was obtained.

3.4. Reference transcription

The reference transcription of the speech material was made by two phonetically trained listeners who had experience in transcribing speech. They transcribed together from scratch and had to agree on each symbol included in the transcript (consensus transcription). They used the CGN symbol set, which is an adaptation of the SAMPA set for Dutch.

3.5. Automatically generated transcription (AGT)

The AGT used as the starting point in this experiment is a simple concatenation of the canonical phonetic representations obtained from the CGN lexicon through a simple lookup procedure. The gaps that remained were filled by obtaining a phonetic transcription from one of a number of sources (Celex English database, Onomastica, etc.).

In these phonetic representations all so-called obligatory word internal processes [5] have been applied, whereas optional word internal processes have not been applied. The phonetic representations obtained from different sources appear to vary with respect to the application of the process of /n/-deletion after schwa [5]. With the sole exception of the degemination process [5], in this concatenated AGT cross-word processes have not been applied.

3.6. Transcription alignment

The AGT was aligned with the Tref by using the Align program [3], in which the distance between corresponding phonemes is calculated on the basis of articulatory features like place and manner of articulation, voice, lip rounding, length, etc. For example, substituting a /t/ for a /d/ has a lower cost than substituting a /t/ for a /x/. In addition to computing insertions, deletions and substitutions, we will also analyse the nature of the discrepancies between the Tref and the AGT.

4. Results

In this section the results of the alignment between the AGT and Tref are examined both from a quantitative (4.1) and from a qualitative point of view (4.2).

4.1. Quantitative results

In Table 2 the frequency of substitutions, deletions and insertions is displayed. The figures express the number and the percentages of symbols that were substituted, deleted and inserted in the AGT when we compare it to the Tref. These data indicate that the quality of this initial AGT is already reasonable if we consider that data on agreement between human transcribers vary between 93.1% and 94.4% for careful speech [10] and between 78.8% and 86.2% for less careful speech [6]. As this AGT is a simple concatenation of canonical forms in which no cross-word processes are represented, the first thing to do is to find out how many of the processes presented in Table 2 take place at word boundaries.

Table 2 Deviations per speech style

category	substitutions		deletions		insertions	
	#	%	#	%	#	%
RS	194	6.4	58	1.9	127	4.2
LC	281	8.3	115	3.4	240	7.1
IN	135	7.3	74	4.0	162	8.8
SC	181	9.4	49	2.5	239	12.4

Table 3 shows the absolute and relative occurrence of substitutions, deletions and insertions at word boundaries for each category of speech style.

Table 3 Substitutions, deletions and insertions at word boundaries

category	substitutions word boundary		deletions word boundary		insertions word boundary	
	#	%	#	%	#	%
RS	95	49.0	45	77.6	53	41.7
LC	164	58.4	93	80.9	111	46.3
IN	72	53.3	53	71.6	102	62.9
SC	107	59.1	34	69.4	129	54.0

The results in Table 3 reveal that roughly half of the substitutions and insertions take place at word boundaries, while for deletions this percentage is even higher. In the following section we further explore the nature of these processes in order to see whether and how the AGT can be improved. While this kind of qualitative analysis can be made both for word-internal and cross-word processes, in this paper we will limit ourselves to cross-word processes.

4.2. Qualitative results

The high frequency of substitutions at word boundaries is worth being investigated in more detail, because from the literature we know that there are various assimilation processes that can take place across words, such as progressive and regressive voice assimilation and nasal assimilation [5]. If the word-boundary substitutions between the AGT and Tref are indeed related to these processes, then it is possible to improve the AGT by modelling this kind of variation. A closer inspection of the data reveals that the majority of word-boundary substitutions can be attributed to differences in the feature voice. Table 4 shows the frequency (column 2 gives the absolute frequencies and column 3 the percentages) of word-boundary voice substitutions per speech type.

Table 4 Word boundary voice substitutions and relative frequency of voice assimilation

category	voice substitutions		Frel
	#	%	%
RS	63	66.3	88.7
LC	78	47.6	86.7
IN	36	50.0	94.7
SC	53	49.5	92.9

Column 4 indicates the relative frequency (Frel) of the cross-word voice assimilation processes in Tref. Frel is calculated by dividing the number of times a process is applied by the number of times the process could have been applied because the conditions for application were met

The data in Table 4 indicate that word-boundary voice substitutions are relatively frequent and that the processes of cross-word voice assimilation are also frequently applied in all four speech types. Such high values of Frel suggest that if the other variant (in this case the one with voice assimilation applied) were chosen, the AGT would approach Tref more closely. Moreover, the relative frequency of cross-word voice substitutions indicates that this kind of static variation modelling (just choosing the most frequent variant) is likely to improve the AGT to a considerable extent. For instance, on the basis of the values presented above the percentages of word-boundary voice substitutions could be reduced to 4.6% (RS), 6.4% (LC), 5.5% (IN), and 6.8% (SC).

With respect to insertions and deletions at word boundaries, a distinction can be made between word initial and word final position. The number of insertions and deletions is slightly higher in word final than in word initial position. Moreover, while the insertions and deletions in word initial position appear to be equally distributed over all phonemes, those in word final position involve a limited number of phonemes. The existence of clear patterns is interesting for us because it suggests the possibility of modelling the phenomena in question with the potential of improving the AGT. Table 5 shows the number of deletions and insertions in word final position and the phonemes involved.

Table 5 Most common deletions and insertions in word final position

	RS	LC	IN	SC
Del w-end	29 66.4%	75 80.6%	41 77.4%	25 73.5%
/n/	23	8	3	6
/m/	0	12	11	3
/@/	1	49	25	11
Ins w-end	33 62.3%	71 63.9%	71 69.6%	79 61.2%
/n/	17	14	23	23
/r/	4	19	22	18
/t/	4	15	12	22

The deletions associated with /n/ appear to be related to the process of /n/-deletion after schwa [5]. Remember that some variants contained /n/ after schwa and others did not. In certain cases, especially in read speech, /n/ was realised after schwa (which means it was present in Tref), which appears from the number of deletions concerning /n/. In other cases, however, /n/ was not realised, as can be inferred from the insertion data. On the basis of these data it is not clear whether /n/ should be present after schwa in the AGT, because any solution would cause some improvements, but probably also an equal number of deteriorations, which means that in this case no strong point can be made for altering the representation in the AGT.

The other types of deletion, concerning schwa and /m/, were caused by the omission of some filled pauses (*uhms* and *uhs*) in the orthographic transcription. In this case the quality of the AGT could be enhanced by improving the orthographic transcription in the first place.

The insertions concerning /r/ and /t/ appear to be related to phenomena known as /t/-deletion and /r/-deletion [5, 7]. It seems that a considerable number of insertions could be avoided by allowing /t/ and /r/ to be deleted in word final position in LC, IN, and SC. Computations of Frel for these processes in Tref revealed that for /r/-deletion Frel varies between 38% in LC and 61.1% in IN. These are typically values which will lead to equal numbers of improvements and deteriorations. For /t/-deletion Frel varies between 12.3% for LC and 20.7% for SC, which are values that do not justify changes in the AGT.

5. Discussion

In this paper we have reported on an experiment that was aimed at comparing an AGT with a reference transcription for the purpose of determining how and in what respects the AGT deviates from Tref, and to what extent the AGT could be improved to make it more similar to Tref. Both the quantitative and the qualitative analyses have provided useful information. In particular, we have learned that the AGT could be improved considerably by incorporating cross-word voice assimilation (both regressive and progressive), since this would dramatically reduce the number of substitutions. With respect to deletions and insertions, on the other hand, the results are less clear-cut. We have found that the AGT could be indirectly amended by improving the orthographic transcription of filled pauses. However, there are no indications that the AGT can be improved by modelling processes such as /t/-deletion, /r/-deletion and /n/-deletion after schwa by selecting one single variant, e.g. the most frequent one. The reason is that, given the values of Frel for these processes, any choice of specific variants will produce both improvements and deteriorations.

The AGT for these processes can be improved further by allowing multiple pronunciation variants. By means of so-called forced recognition, it is possible to let a continuous speech recognizer (CSR) decide which of different pronunciation variants of one and the same word best fits the acoustic signal [8]. A CSR in forced recognition mode does not have to select words from its entire lexicon, but from a limited subset of variants of one word, for instance one with /n/ and one without /n/. A prerequisite for applying this procedure is that the orthographic transcription of the utterance is available. Since in the CGN context all material is orthographically transcribed, forced recogniton can be applied. This method has been shown to select variants with levels of accuracy that are comparable to those of human listeners [8]. Forced recognition could turn out to be fruitful also for those processes that were explicitly not considered in this paper, such as the non-voice related cross-word substitutions, the word initial deletions and insertions and all word internal substitutions, deletions and insertions.

6. Conclusions

The results of the experiment reported on in this paper have provided insight into the nature of the deviations between an AGT and a Tref. For some speech types, these deviations appear to be in the order of magnitude of the deviations between human transcribers. Moreover, some deviations exhibit clear patterns which are related to well-known phonological processes. This suggests that there are possibilities of improving the quality of the AGT by modelling the phonological processes observed, thus obtaining an AGT that should be more similar to Tref than it is now.

Considerable improvements could already be obtained through the simple procedure of choosing the most frequent variant. An AGT of even better quality could be obtained by having an ASR in forced recognition mode choose the variant that best matches the speech signal.

After the AGT has been improved according to this procedure, one could analyze the discrepancies that still remain between the AGT and Tref and decide on the following course of action for different types of speech: taking the AGT as is or having human transcribers correct it. In both cases greater efficiency is guaranteed than with the initial AGT or with human transcriptions made from scratch.

With respect to the more general context in which this experiment was carried out, that of the CGN project, we may conclude that the idea of exploring the potential of AGTs was indeed worth investigating, since this study has produced several important findings. First, that even a simple concatenation AGT can be of reasonable quality for certain speech types. Second, that such an AGT can be further improved. Third, how this AGT can be improved.

To conclude, our recommendation to those who have to organize phonetic transcriptions of huge amounts of speech data would be to consider the possibility of optimizing automatic transcriptions before embarking on costly enterprises with human transcribers which do not necessarily guarantee higher quality transcriptions.

7. References

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

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

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

[4] Shriberg, L.D., Kwiatkowski, J., and Hoffman, K. "A Procedure for Phonetic Transcription by Consensus". *Journal of Speech and Hearing Research*, **27**, 456-465, 1984.

[5] Booij, G. *The phonology of Dutch*, Clarendon Press, Oxford, 1995.

[6] Kipp, A., Wesenick, B., and Schiel F. Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Proceedings EUROSPEECH '97*, 1023-1026, 1997.

[7] Cucchiarini, C., and Van den Heuvel, H.. Postvocalic /r/-deletion in Dutch: more experimental evidence. *Proceedings ICPhS San Francisco*, 1673-1676, 1999.

[8] Wester, M. Kessens, J.M. Cucchiarini, C. and Strik H. Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer, to appear in *Language and Speech,* 2001.

[9] Hoste, V., W. Daelemans, E. Tjong Kim Sang, and S. Gillis. "Meta-learning for phonemic annotation of corpora" In *Proceedings of ICML-2000*, P. Langley (ed), 375-382. Stanford University, 2000.

[10] Kipp, A., Wesenick, B., and Schiel F. Automatic detection and segmentation of pronunciation variants in German speech corpora, *Proceedings ICSLP '96*, 106-109, 1996.

8. Acknowledgements

This research was supported by the project "Spoken Dutch Corpus (CGN)", which is funded by the Netherlands Organization for Scientific Research (NWO) and the Flemish Government. We are indebted to the members of CGN signal processing group for their comments on an earlier version of this paper.