PHONETIC TRANSCRIPTIONS OF LARGE SPEECH CORPORA

Cover design by Jeroen Nijboer and Caspar Ong – LogicaCMG, Amstelveen

The cover design refers to the movie The Matrix (1999). Not only was the calculation of the distance between two transcriptions based on figures organised in a matrix, but more essentially, the observations reported in this thesis are also based on the assumption of the existence of an absolute truth that can only be approximated.

Printed and bound by PrintPartners Ipskamp, Nijmegen

ISBN-10: 90-9020394-X
ISBN-13: 978-90-9020394-2

PHONETIC TRANSCRIPTIONS OF LARGE SPEECH CORPORA

een wetenschappelijke proeve op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus prof. dr. C.W.P.M. Blom
volgens besluit van het College van Decanen
in het openbaar te verdedigen op vrijdag 7 april 2006
des namiddags om 1.30 uur precies

door
Diana Maria Binnenpoorte
geboren op 19 februari 1974
te Nijmegen

Promotor	Prof. dr. L. Boves
Co-promotor	Dr. C. Cucchiarini
Manuscriptcommissie	Prof. dr. R. van Hout (Voorzitter)
	Prof. dr. A. Braun (Philipps-Universität Marburg, Duitsland)
	Dr. J. Duchateau (Katholieke Universiteit Leuven, België)

DANKWOORD

Het is gelukt. Er is een berg bedwongen, een berg die eerst van een afstand helemaal zo hoog niet leek, maar die, naarmate het einde in zicht kwam, steeds zwaarder begaanbaar werd. Toen ik in 1999 op de afdeling Taal & Spraak kwam werken, had ik niet te bedoeling te promoveren. Ik ben nu ontzettend blij dat ik uiteindelijk toch op het voorstel van m'n promotor, Loe Boves, en co-promotor, Catia Cucchiarini, ben ingegaan, en daadwerkelijk dit proefschrift heb geschreven over een deel van het onderzoek dat ik de afgelopen jaren heb uitgevoerd. Het moge duidelijk zijn dat ik deze twee mensen verschrikkelijk dankbaar ben en naast hen nog veel anderen. Daarom zal deze, waarschijnlijk meest gelezen, sectie van het proefschrift helemaal gewijd worden aan iedereen die me geholpen heeft.

Zoals gezegd wil ik mijn promotor, Loe, bedanken voor het vertrouwen dat hij uitsprak toen hij voorstelde om een proefschrift te schrijven ten tijde van het CGN project. Er zijn momenten geweest dat zijn vertrouwen in een goede afloop groter was dan dat van mij, waardoor ik nooit heb willen opgeven. Daarnaast wil ik Loe bedanken voor alle praktische hulp, vooral toen er aan het voltooien van dit proefschrift maar geen einde leek te komen.

De samenwerking met Catia is eigenlijk al begonnen toen het CGN project net gestart was en we beiden 'iets moesten uitzoeken' voor de fonetische transcripties. Vanaf dat moment heeft ze me met haar enthousiasme en vindingrijkheid laten zien hoe leuk het is om onderzoek te doen. Catia wil ik ook bijzonder bedanken voor het geduld waarmee ze me geholpen heeft om papers en artikelen te schrijven.

Het CGN project is al een paar keer genoemd en van Nelleke, de meest betrokken projectleider die ik ken, heb ik geleerd dat er geen problemen bestaan zonder uitvoerbare oplossingen. Daarom wil ik haar bedanken voor haar belangstelling en hulp, niet alleen voor mij of mijn werk, maar voor iedereen op de afdeling. Alle andere CGN-ers, met name de Belgische vrienden en het Nijmeegse team wil ik bedanken voor de goede sfeer waarin we met ons allen zo hard gewerkt hebben. En vooral Andrea, mijn kamergenoot in 17.18 - de zoete inval - wil ik in het bijzonder even noemen. We hebben wat afge-cgn-d! En hadden onze pc's een werkende ASR met APT gehad, dan waren we nu een flinke collectie dvd's rijker met ons eigen CGW (Corpus Gesproken -lange- Weekendverhalen), CGR (CGRoddels – alleen waarheidsgetrouwe), CGL (CGLevenswijsheden – ahum) en CGF (CGFrustraties – grote en kleine). Deze corpora zouden wellicht niet representatief zijn voor het hedendaags Nederlands, maar zeker wel voor hedendaagse vriendschap! Daarnaast wil ik Simo, de Utrechtse dependance, bedanken voor de expert transcripties en de fijne samenwerking.

Voor het CGN project heb ik ook een groot deel van de tijd bij Taal & Spraak bij SPEX gewerkt en als Expexer wil ik de Spexies dan ook bedanken en vooral Eric omdat hij me belde in 1999 met de vraag wanneer ik kon beginnen. Alle (oud-) collega's van Taal & Spraak moet ik hier eigenlijk noemen, vanwege de vriendschappelijke en goede sfeer, waarvan ik nu weet dat die uniek is, en alle onbaatzuchtige hulp bij het voorbereiden van presentaties en posters, bij het geven van opmerkingen en suggesties voor papers en delen van dit proefschrift, bij het snijden van taarten, en gewoon bij de dagelijkse werkzaamheden. Maar in het bijzonder noem ik Janienke omdat ik altijd bij haar terrecht kon en zij aan een half woord genoeg had om te begrijpen wat ik bedoelde, Christophe vanwege de inhoudelijke discussies en zijn geestdrift, en Hella voor haar bemoedigende woorden.

Werken in Nijmegen betekende voor mij natuurlijk ook treinreizen. De trainladies (Janienke, Dorota, Andrea, Judith, Simo en Febe) maakten die 2 x 50 minuten vol evaluerende werkbesprekingen, puzzels, kranten en kamelen tot een leuke tijd waar geen vertraging tegenop kan. Toen dit proefschrift eindelijk (bijna) af was en ik kon gaan nadenken over paranimfen, wist ik eigenlijk al direct wie ik zou vragen: Andrea en Janienke (let op je dankwoordindexcijfer) en ik wil hen dan ook bedanken omdat ze toen zo enthousiast 'ja' riepen en dat nu nog steeds zijn.

Uiteraard wil ik ook mijn (schoon-) familie en lieve vrienden bedanken voor het feit dat ik nooit heb hoeven uitleggen wat ik nou precies deed en vooral waarom en dat zij toch altijd zonder twijfel hun steun hebben uitgesproken. En als laatste wil ik natuurlijk Ferdinand ontzettend bedanken. Niet alleen voor het layouten van het manuscript terwijl ik zenuwachtig op en neer drentelde, maar juist voor al zijn begrip, steun, liefde en zijn zo belangrijke relativeringsvermogen. Iedereen, enorm bedankt, de andere kant van de berg ziet er prachtig uit!

CONTENTS

INTRODUCTION

CHAPTER 1

1.1 Introduction

1.1.1 Real-life speech

Speech is the most basic communication medium in which language is involved. Everyday spoken communication between language users is based on the ability of speakers to produce intelligible speech, and, at the same time, on the ability of listeners to understand the spoken message of the speaker. The latter is actually quite amazing given the fact that each time a word is uttered, its pronunciation differs. This variation does not only occur between speakers – *inter-speaker* variation – but also in the speech of one and the same speaker – *intra-speaker* variation (Woodland, 1998). Inter-speaker variation is caused by differences between speakers, including differences in age, gender, accent due to differences of the region where the speakers went to school, and other social factors such as educational level (Laver, 1994; Scherer & Giles, 1979). Intra-speaker variation is caused by factors such as conversational setting or genre, the topic, the level of formality, the emotional state of the speaker, the interlocutor, and also effects of running speech, for instance assimilation of speech sounds (Polzin & Waibel, 1998; Weintraub et al., 1996). All these factors co-occur and influence the actual pronunciation.

The degree of pronunciation variation that may occur in an informal spontaneous conversational setting is illustrated in Figure 1-1. The example is extracted from the Spoken Dutch Corpus (CGN) (Boves & Oostdijk, 2003; Oostdijk, 2000). The ORT-line contains the verbatim transcription of the conversation. The FON-line is a manually generated broad phonetic transcription representing how the words were actually pronounced. The CAN-line gives the canonical transcription, which is a representation of the individual words as they appear in a pronunciation dictionary. The fragment was taken from an ongoing conversation between two friends discussing a soap opera. The recording was made in a living room. The same conversation comprises numerous other passages that demonstrate the same degree of variation as in the selected example. A quick comparison between the manual phonetic transcription and the canonical representation of the orthographic words shows that in a real-life situation the pronunciation of some words is rather different from the dictionary pronunciation. Speaker N01152 seems to be permitted to omit complete syllables and still speaker N01151 seems to understand what has been said (given the confirmative ‘yes’). Clearly, there is no one-to-one relation between the orthography or the canonical representation and the actual pronunciation. In fact, the example shows that the distance between the orthography and the pronunciation here is

rather large and diverse (see e.g. two instances of the word ‘eigenlijk’ in Figure 1-1), which makes it difficult to predict actual pronunciation from orthography alone.

N01151:	ORT	ja ja. *(yes yes.)*
	FON	ja ja
	CAN	ja ja
N01152:	ORT	je had eigenlijk ik uh... *(there was actually I eh...)*
	FON	j@ hAt E+Nk ik @
	CAN	j@ hAt E+G@l@k Ik @
N01152:	ORT	je had iedere keer zo’n deelopnames *(each time there were such partial recordings)*
	FON	j@ hAt id@ k@ son delOpnam@s
	CAN	j@ hAt Id@r@ ker zon delOpnam@s
N01152:	ORT	terwijl je eigenlijk het geheel ook nog 'ns wou zien. *(while you actually wanted to see the whole once again too.)*
	FON	t@wE+l j@ E+k @t x@hel ok nOG @s wA+ zin
	CAN	tErwE+l j@ E+G@l@k hEt x@hel ok nOx @ns wA+ zin
N00151:	ORT	ja. *(yes.)*
	FON	ja
	CAN	ja

Figure 1-1 Transcription of a piece of fragment fn000771, time interval of conversation 190.18 – 194.49 s. Two speakers (N01152 and N01151) were recorded during a face-to-face conversation in a living room environment. The three lines per utterance are, on top, ORT, the orthography, second, FON, the manual phonetic transcription, and the last line is CAN, the canonical representation.

The example in Figure 1-1 already gives an impression of the wide range of possible pronunciations. If one is interested in research on pronunciation phenomena, many more phonetically transcribed ‘examples’ of real-life speech need to be collected. Such large collections of transcribed speech recordings are referred to as *speech corpora*. Large speech corpora constitute a rich resource for empirical investigations on spoken language. But before corpora are usable as tools for pronunciation research, it is necessary to obtain a *phonetic transcription* of the speech material.

1.1.2 Speech corpora

Most of the speech research carried out until the 1990s concerned carefully pronounced speech, often recorded in a laboratory setting. The absence of large storage devices and automatic techniques for processing large amounts of speech forced researchers in those days to confine themselves to small amounts of speech data. Owing to this, their research was mainly focussed on speech of a limited number of speakers. In the eighties the automatic speech recognition (ASR) research community required speech corpora to be used for the training of statistical acoustic models (Lamel et al., 1986; Price et al., 1988). The first large multi-speaker corpora, Macrophone (Bernstein et al., 1994) and the Dutch Polyphone (Den Os et al., 1995), were typically recordings of prompted speech, such as reading aloud sentences, lists of words, etc., recorded over the telephone. The applications in which the speaker-independent ASR systems of the mid-nineties could play a role were mainly short command voice response systems (see also Van den Heuvel et al., 2001).

As the ASR systems improved over the years (Rudnicky et al., 1994), the type of speech-driven applications changed towards more natural dialogue systems. At the same time, the availability of large digital storage devices paved the way for the compilation of larger speech corpora, which, in addition, contained considerable portions of natural (spontaneous) speech, for example Switchboard (Godfrey et al., 1992), and Verbmobil (Hess et al., 1995).

Besides the inclusion of large spontaneous components in the newer speech corpora, many present-day corpora are designed to serve as multi-purpose resources, viz. suitable for more types of linguistic research than developing ASR systems alone. Examples of these multi-purpose corpora are JSC (Furui et al., 2000) for Japanese, SALAR (Wissing et al., 2004) for South-African languages and CGN (Oostdijk, 2000) for Dutch. These corpora contain speech recorded in several situations (lessons, conversations), through various channels (radio, telephone, headset) and are enriched with multiple annotation layers, such as a phonetic transcription, Part-Of-Speech tags, prosodic and syntactic annotation, besides the conventional orthographic transcription.

For a large collection of speech recordings to be referred to as a genuine speech corpus, the speech material needs at least an orthographic transcription. An orthographic transcription is a verbatim representation of the speech using standard spelling conventions. Such an orthographic transcription is a first indication of *what* was actually spoken in the speech recordings. In many languages there is a substantial distance between the sound and the spelling of words (Wells, 1996), as was also illustrated in the example of spoken Dutch in Figure 1-1. If one is interested in *how* the speech sounds were realised, a (broad or narrow) phonetic transcription is indispensable.

In the early days, when research in ASR addressed read speech, an orthographic transcription with a pronunciation dictionary containing the canonical phonetic transcriptions was considered sufficient. As the type of speech to be processed became more and more spontaneous, a simple canonical representation was no longer sufficient and phonetic transcriptions became necessary. Moreover, other research areas focussing on spontaneous speech effects, such as phonetics, phonology, and sociolinguistics, also require phonetic transcriptions of the speech recordings in a corpus.

1.2 Phonetic transcriptions of large speech corpora

As suggested above, phonetic transcriptions of large speech corpora constitute a very useful and rich resource for linguistic research. It is self-evident that, before phonetic transcriptions can be used, they need to be generated. That this is not a straightforward task will be explained in the following sections. The main focus of this thesis is on the generation of phonetic transcriptions of large speech corpora, and, in relation to this, on the evaluation of the quality of these transcriptions.

In the past, a phonetician or speech researcher generated phonetic transcriptions of the speech material he or she was working on by hand. However, since the possibilities of sound storage devices have grown, larger amounts of speech recordings can now be collected, making the generation of phonetic transcriptions by hand problematic. A complete manual phonetic transcription of the speech recordings in an average-sized, present-day speech corpus made by an expert phonetician is practically impossible, if not for budgetary and time-related reasons, then for the difficulty to find trained phoneticians who are willing to carry out such a tedious task. To illustrate, a broad phonetic transcription of one minute of conversational speech takes about sixty minutes when generated from scratch.

Alternative procedures need to be developed and deployed to create phonetic transcriptions of large speech corpora. For instance, one can decide to manually transcribe only parts of a whole corpus, or, as was actually the case in many corpora that were recently compiled, to employ (semi-)automatic techniques in the transcription procedure (Furui et al., 2000; Godfrey et al, 1992; Hess et al., 1995; Pitt et al., 2005; Wissing et al., 2004). These (semi-)automatic procedures are primarily intended to minimise human labour as much as possible. However, if it appears that the quality of automatic transcriptions is good enough, involving human labour can be completely avoided. Automatically generated phonetic transcriptions have a substantial cost advantage over manually generated transcriptions. In addition, automatic transcriptions are more consistent, and certainly more reproducible than transcriptions of individual experts. After

all, human-made transcriptions are known to contain elements of subjectivity and idiosyncrasies (Cucchiarini, 1993).

The success of an automatic phonetic transcription procedure for large speech corpora that should minimise or even replace human transcription labour depends on the ultimate quality of the resulting transcription. An automatic phonetic transcription can be considered of good quality if its quality is comparable to that of a human-made transcription. On top of this, the basic purpose of phonetic transcriptions, either manually or automatically generated, is to represent the actual pronunciation of the speech. Consequently, it is essential to determine whether the phonetic transcriptions of large speech corpora indeed fulfil that requirement. Perhaps somewhat surprisingly, there are no generally accepted and efficient methods for measuring transcription quality. Therefore, part of the research reported in this thesis was devoted to developing a reliable and objective quality measure.

1.3 The issues

Several issues regarding the generation and evaluation of phonetic transcriptions of large speech corpora are addressed in this thesis. In this section these issues will be discussed on the basis of the following two questions.

1. How to get from orthography to phonetic transcription
2. How to measure phonetic transcription quality

1.3.1 From orthography to phonetic transcriptions

Grapheme-based transcriptions

The most elementary type of automatic phonetic transcription is based on the orthographic transcription alone. In this procedure it is assumed that all relevant information is contained in the orthographic transcription, and there is no need to consult the speech signal again. The phonetic representations of the orthographic words are automatically obtained through a lexicon look-up procedure. For this method a pronunciation dictionary is consulted that contains canonical representations of orthographic words. In case orthographic words are missing from the dictionary, grapheme-to-phoneme conversion can be used as a fall-back option. Once the canonical representations are collected, these forms are concatenated to yield a phonetic transcription of the input text. As said before, this approach has proved successful for ASR research addressing carefully produced speech.

However, this method comes with a number of fundamental problems. The first problem that this grapheme-based method encounters is based on the fact that many words,

at least in the English and Dutch language, have more than one generally accepted pronunciation according to pronunciation dictionaries. For example the Dutch words 'politie' ('*police*') and 'Nobelstraat' ('*Nobelstreet*') can have the following entries /politsi/ and /polisi/, and /nob@lstrat/ and /nobElstrat/, respectively. No matter which of the alternative pronunciations is taken from the dictionary, there is always the risk that it was the wrong one. The second problem is a fundamental shortcoming of transcribing without consulting the speech signal. In real-life conversational speech large discrepancies can exist between the actual pronunciation and the canonical representations of the orthographic words. This was illustrated by the example in Figure 1-1. The suitability of the conversion methods as transcription procedure is therefore dependent on the degree of deviation between the actual pronunciations and the canonical, or dictionary, representations.

The quality of grapheme-based transcriptions can possibly be improved by applying some phonological rules in a post-processing phase in order to model some highly frequent cross-word pronunciation processes. The issue that arises then is how to obtain these rules in the first place, and, how to determine which rules are applicable where and when.

Two knowledge sources can be distinguished in this respect, the literature, and already existing corpora. Literature as a knowledge source may not provide a satisfactory solution, since most studies reported in the literature are based on laboratory and read speech only, while most newly compiled corpora contain a substantial amount of spontaneous speech. The other option, extracting phonological knowledge from existing speech corpora that contain a phonetic transcription, can of course only be used if these corpora are available and contain relevant information, viz. transcriptions of similar speech styles for which transcriptions need to be generated. In a situation in which neither source is available for phonological knowledge extraction, other solutions need to be found.

ASR as transcription tool

In contrast to the method mentioned above, the automatic transcription methods discussed in this section actually do take the speech signal into account. Two transcription techniques are briefly discussed: 1. free phone recognition; 2. forced recognition. For both methods an Automatic Speech Recogniser (ASR) is needed.

The first technique, free phone recognition, follows the same principles as normal ASR (see Wester (2002) for a brief description) with the main difference that instead of words, phonemes have to be recognised. The recognition lexicon contains a list of phonemes, instead of a list of words. The recogniser is often constrained by 'language' models that describe the phonotactic constraints of the language and other tuning parameters in order to improve the performance.

The second technique, forced recognition, uses the orthographic transcription that is already available in large speech corpora. In a recognition lexicon the canonical representations for each orthographic word are contained. For each word plausible pronunciation variants are generated and included in the recognition lexicon. The task of the ASR is to choose the pronunciation variant that best matches the speech signal such that a new transcription is obtained (Kessens, 2002; Strik & Cucchiarini, 1999). This technique is employed by many researchers (see Strik & Cucchiarini (1999) and references therein) as an automatic phonetic transcription technique. The main issue is how to obtain plausible pronunciation variants, or more fundamentally, how to obtain knowledge in order to generate the variants.

As said before, both the literature and existing speech corpora are potential knowledge sources. Assuming that relevant literature is available, a knowledge-based approach can be followed to generate plausible pronunciation variants. Pronunciation rules that are formulated based on the phonological knowledge described in the literature can be used to generate pronunciation variants to be included in the recognition lexicon. A similar procedure can also be followed in a data-driven approach, when pronunciation rules are extracted from a hand-transcribed speech corpus. For that goal a decision-tree tool can be used to derive pronunciation rules based on the alignment between the transcription that represents the actual pronunciation and the canonical forms (Riley & Ljolje, 1996). However, applying rules to generate pronunciation variants has its limitations. For instance, it is likely that extreme reductions of frequently occurring words in spontaneous speech are difficult to capture into rewrite rules. The example in Figure 1-1 shows substantial deviations between the actual pronunciation of the word ‘eigenlijk’ (‘actually’) and the canonical representation. Moreover, these deviations differ between different realisations of the word, even by the same speaker in the same conversation. This makes it virtually impossible to formulate comprehensive and accurate rules for predicting plausible variants. When the rules needed to map the canonical representation of ‘eigenlijk’ onto the observed pronunciations are applied to other words, they will most probably result in forms that are inappropriate. At the same time it is clear that using only general rules will not produce the observed reduced forms. Therefore, other methods besides rule-based variant generation are necessary to obtain pronunciation variants that have to be included in the recognition lexicon. One possibility is adding forms that have been observed in a transcribed corpus (cf. chapter 3).

For developing phonetic and linguistic knowledge, it is not enough to simply collect pronunciation variants; it is also necessary to systematically describe where and when these variants are permitted in spoken language. This is especially true for extremely reduced variants. For example, it may be that some reduced forms are restricted to specific contexts,

e.g. when a word occurs as part of a frequent expression. It is obvious that an existing speech corpus containing relevant material constitutes a rich resource for investigating pronunciation phenomena that are either not yet described in the literature, or are not easily covered by rewrite pronunciation rules.

Manual transcriptions

Despite the practical and financial problems alluded to above, another method to obtain phonetic transcriptions is to take recourse to human transcribers. As argued above, a complete manual phonetic transcription of a large speech corpus made from scratch is practically impossible. The main problem in this respect concerns the costs. A widely used solution to speed up the process and thereby reduce the costs is an approach in which human transcribers verify and correct a, preferably, optimised automatically generated transcription. This approach leads to a considerable reduction in time compared to transcribing from scratch.

However, the verification and correction method introduces new problems of its own. First, in case large amounts of speech data need to be transcribed the verifications are inevitably carried out by several human transcribers. The fact that human transcriptions contain an element of subjectivity and idiosyncrasies (Cucchiarini, 1993) increases the probability of inconsistencies in the ultimate transcription. Second, making auditive transcriptions remains a difficult task, even when this is achieved through verifying and correcting an example transcription. The more deviant the pronunciations from the canonical form, the more complex the task for the human transcriber to identify the various speech sounds. Furthermore, the transcriptions have to be made under time-pressure, given the amount of data that needs to be transcribed, which complicates the task even more. Third, using an example transcription can increase the chance of biasing the resulting transcription towards the given example transcription.

In spite of the fact that a verification and correction method leads to cost reduction compared to transcriptions made from scratch, the above mentioned problems and risks raise questions about the added value of having human transcribers correct an automatically generated phonetic transcription.

1.3.2 Measuring transcription quality

After having addressed the issues concerning the generation processes of phonetic transcription, the second step is to establish the quality of the transcriptions. Phonetic transcription quality expresses the extent to which the string of symbols is a valid representation of what was actually realised. The most straightforward manner to measure

the quality of a transcription would undoubtedly be by comparing the symbols of the transcription in question with the symbols of a transcription that represents the ground truth: a reference transcription. However, such a unique and true reference transcription does not exist (Cucchiarini, 1993).

The problem that arises then is how to approximate such a reference transcription in order to measure transcription quality. Individual human expert transcriptions are known to be subjective and to contain idiosyncratic elements, which makes them unsuitable as a reference transcription. Therefore, measures should be taken to minimise these subjective and idiosyncratic elements.

An attractive solution is a consensus transcription, as suggested by Shriberg et al. (1984), obtained through a procedure in which a group of transcribers discuss and negotiate to reach agreement on each single symbol in the transcript. By following such a procedure, idiosyncrasies, errors, and subjective impressions in the transcriptions are likely to be reduced substantially, which makes a consensus transcription a suitable approximation of the unique and true reference transcription.

Another possibility to establish the quality of human-made transcriptions would be to determine inter-transcriber agreement, i.e., the degree of agreement between transcriptions of the same material made by several transcribers. High agreement scores, or high consistency between transcriptions, may indicate that the transcriptions are valid representations of the actual speech signal. After all, if a speech signal is transcribed with the same symbols by hundred different transcribers, it is very likely that the symbols are a correct representation of the speech signal. At least two issues complicate the usage of inter-transcriber agreement as a sole quality measure in the context of large speech corpora. First, the requirement of repeated measurements can often not be met because in large speech corpus projects the number of transcribers is sometimes quite limited, and it is unusual to have multiple transcriptions of (part of) the speech signals. Second, it is often decided to have the human transcribers verify and correct a given example transcription instead of having them transcribe from scratch. In fact, agreement scores of 100% can be obtained if the example transcription is left intact, which of course does not mean that the transcriptions are indeed valid representations of the actual speech signal.

Assessing automatically generated transcriptions

The aim of automatic phonetic transcriptions of large speech corpora, i.e., to minimise or replace human transcription labour, is elusive without the availability of a quality assessment in order to judge whether the aim was realised. First, the quality of the automatic transcription has to be established and, as explained above, the best option is to compare it with a reference transcription. Second, it should be determined if the measured

quality is indeed good enough such that the automatic transcription can replace human-made transcriptions. For this purpose, a procedure is needed that allows one to define a threshold on the quality measure, above which the quality of the automatic transcription can be considered as equivalent to (or perhaps even better than) human-made transcriptions in large speech corpora.

At present, there are no clear-cut criteria upon which such a threshold can be based. Thus, it is necessary to try and make the concept 'comparable to human performance' operational. The obvious way to operationalise this criterion is by stipulating that automatic transcriptions must show the same degree of deviation from human transcriptions as multiple human transcriptions of the same speech signals deviate from each other. The latter measure is conventionally expressed in terms of inter-human agreement. However, since human transcriptions in large speech corpora are often made by editing an example transcription, inter-human agreement might give a too optimistic view. The example transcription presented to the human transcriber is likely to affect the measured agreement between the various transcriptions. If human transcribers leave the example transcription unchanged, for whatever reasons, high agreement scores are the result. This makes the criterion imprecise and therefore no strict threshold can be defined based on straightforward inter-transcriber agreement scores. Moreover, it is very well possible that future applications that make use of phonetic transcriptions of large speech corpora each require different quality levels, and therefore demand other thresholds that need to be achieved. Therefore, in defining a threshold for automatic transcription quality these issues should be kept in mind.

Assessing manually verified transcriptions

The reason for having human transcribers verify and correct a given example transcription instead of transcribing from scratch is to reduce transcription time and therefore costs. However, it is unclear how this procedure affects the quality of the resulting transcriptions.

Inter-transcriber agreement seems to be an obvious measure to establish the quality of human-made transcriptions. However, since the speech recordings in large speech corpora are non-recurring, an additional experiment must be set up in which all transcribers have to transcribe the same speech sample to measure inter-transcriber agreement. Furthermore, the number of observations that can be made to measure inter-transcriber agreement is limited by the number of transcribers actually employed in a large speech corpus project. From the documentation that comes with existing corpora it seems that there are seldom more than ten different linguists (or students) involved in making phonetic transcriptions. On top of this, inter-transcriber agreement as a sole quality measure for transcriptions of large speech corpora has additional limitations. First, it is to be expected that human

transcribers can only concentrate on a limited number of processes and edit the example transcription in this respect, while for the remaining processes the example transcription is left intact. Second, considering the fact that the human transcribers work under a certain time pressure, it is even more likely that the example transcription is left intact too often. If human transcribers leave the example transcription unchanged, high agreement scores are the result, whereas this does not necessarily mean that the transcriptions are indeed good representations of the actual pronunciation. So, an additional objective measure is required that can determine whether the transcriptions do represent the actual pronunciation. This can be achieved by comparing the human-made transcriptions with a close approximation of a true reference transcription, for instance a consensus transcription.

1.4 Goals and approaches

The first goal of the research described in this thesis is to develop procedures to generate broad automatic phonetic transcriptions of large speech corpora. To reach that goal, several problems must be solved. One of these problems concerns the lack of phonological knowledge needed for automatic generation of phonetic transcriptions for unprepared (and therefore probably less canonical) speech. Another problem is the absence of published pronunciation rules needed to create plausible pronunciation variants for spontaneous speech and the limitations of rule-based approaches when it comes to transcribing more spontaneous pronunciations.

The second goal of the research is to develop validation measures, first, to evaluate the automatic transcription procedures by assessing the resulting automatic phonetic transcriptions, and second, to evaluate manually generated phonetic transcriptions. The problems that arise here first concern the definition of an objective quality measure given the apparent unsuitability of inter-transcriber agreement as a sole quality measure when calculated for transcriptions obtained by manual editing an example transcription. Secondly, a threshold has to be established indicating whether the automatic transcriptions are of sufficient quality and can therefore replace human-made transcriptions of large speech corpora. In this section we present a brief description of how these issues are tackled.

1.4.1 Generating phonetic transcriptions

For the automatic generation of phonetic transcriptions of large speech corpora knowledge about phonetic and phonological processes can be employed. Both an optimised grapheme-based transcription and an ASR-based transcription require this type of knowledge to

model pronunciation phenomena in real-life speech. If available, the knowledge can be found in the literature or can be extracted from already phonetically transcribed speech corpora. However, in a situation in which both relevant literature and data are lacking as knowledge sources, other procedures need to be developed. Such a situation is not exceptional; many newly compiled large speech corpora contain a substantial amount of recordings of extemporaneous speech, whereas most of the phonological descriptions in the literature are based on analyses of laboratory and read speech. Other data resources, apart from newly compiled corpora, often do not contain the same type of speech recordings or are simply not available. So, the question that arises is how to obtain automatic phonetic transcriptions for a large speech corpus that contains speech of under-researched speech styles, without ignoring the quality aspects of the automatically generated transcription.

The solution to this problem that we propose consists of a bootstrapping procedure. This procedure is both aimed at improving automatic transcription generation, and at obtaining new systematic knowledge on the nature and frequency of phonological processes occurring in various speech styles. The iterative procedure is designed in such a manner that the newly obtained information can subsequently be deployed to generate novel pronunciation rules for further improvement of automatic phonetic transcriptions. Validation of the automatic transcriptions after each iteration cycle offers the possibility of deciding whether the automatic transcription is of sufficient quality.

The bootstrap procedure sketched in the previous paragraph can start with almost no information and knowledge beyond a verbatim transcription. Alternatively, in a situation in which a large speech corpus that contains phonetic transcriptions is already available, the question arises on how to extract relevant phonological knowledge from such a corpus and how this knowledge can subsequently be employed for generating automatic phonetic transcriptions of a new corpus. Since most available phonological knowledge concerns read speech and the quality levels of automatic transcriptions for spontaneous speech are likely to be way below the quality levels for read speech, we will concentrate on obtaining information about pronunciation processes in spontaneous speech. In using this information we attempt to model some extreme pronunciation phenomena that are often observed in spontaneous speech, see Figure 1-1 in section 1.1.1. Because these extreme pronunciations are difficult to generate by means of rules, an alternative method must be developed. Therefore, we propose and test a procedure in which we extract frequently observed pronunciation variants and collect information on the prior probabilities of these pronunciations from a hand-transcribed corpus of spontaneous speech. Both the list of pronunciation variants and their probabilities are relevant information for an ASR in forced

recognition mode. It is likely that more accurate automatic phonetic transcriptions of spontaneous speech can be generated in this manner.

The availability of a large corpus of phonetically transcribed spontaneous speech additionally offers the possibility of analysing pronunciation phenomena on a different level, i.e., beyond word boundaries. Are there systematic patterns that can predict the occurrence of some of the extreme pronunciations observed in spontaneous speech? More precisely, do words show more peculiar pronunciations when they occur in fixed expressions as opposed to when they occur in any other context? If there are any systematic patterns, this information can be used for better modelling spontaneous speech processes when generating automatic phonetic transcriptions.

In order to answer these questions we have used a large corpus of spontaneous speech from which frequent word sequences have been extracted and analysed according to a set of criteria which relate to the linguistic concept of multi-word expressions. The sequences have been analysed both with respect to their lexical status and the actual pronunciations as observed in the corpus.

1.4.2 Measuring transcription quality

The approaches suggested above are all aimed at improving the quality of automatically generated transcriptions by obtaining more relevant knowledge for modelling speech processes. However, the question as to when an automatic transcription is good enough to replace human transcription effort remains to be addressed. As argued in section 1.3.2, the most objective quality measure would be based on a comparison between the automatic transcription and a 'true' reference transcription. As the latter does not exist, we try to approximate the true reference transcription by having a sample of the speech material transcribed by two expert transcribers in consensus mode and then compare the automatically generated transcription with the reference transcription of the same sample. Once the degree of agreement between the automatic transcription and the reference transcription is established, a threshold must be defined to determine if the quality of the automatic transcriptions is good enough to replace human-made transcriptions of large speech corpora. This threshold should preferably be based on human transcribers' performance as measured in their transcriptions of large speech corpora. Due to the absence of reference material, i.e., human-made transcriptions of the same type of speech, we must resort to inter-human agreement scores reported in the literature. In setting thresholds we must take into consideration both the fact that inter-transcriber agreement scores obtained for transcriptions that were made by editing an example transcription are

likely to be inflated and the fact that automatic transcriptions have a considerable cost advantage over human-made transcriptions.

In order to reduce transcription time and therefore money human-made transcriptions in large speech corpora are usually produced by following a procedure in which several transcribers edit an example transcription. The issues here are to what extent such a procedure is reliable. What is the added value of humans editing an automatically generated example transcription? Are these human-made transcriptions consistent? Are the transcriptions affected by a bias towards the example transcription? These questions can be answered by setting up an experiment in which several transcribers transcribe the same speech sample by editing an example transcription. To measure transcription quality we adopt various measures, viz., inter-transcriber agreement and a comparison between the human-made transcriptions and a consensus transcription. Besides these quantitative measures a qualitative analysis is made that reveals the nature of the discrepancies between various human-made transcriptions and the reference transcription.

1.5 Material: The Spoken Dutch Corpus – CGN

The research described in this thesis is carried out on a large, multi-purpose corpus of spoken Dutch, the Spoken Dutch Corpus, *Corpus Gesproken Nederlands* (CGN for short). Since this corpus plays an essential role in the following chapters, in this section, the CGN, its design, and the transcription procedures are explained.

Between 1998 and 2004 the Spoken Dutch Corpus was constructed. The final release in March 2004 contains about 9 million words of contemporary Dutch as spoken by adults in Flanders (one-third of the material) and the Netherlands (two-thirds of the material). The main reason for the compilation of this multi-purpose corpus was to satisfy the need of linguists and speech technologists and researchers for a large resource of spoken Dutch. Before the CGN was released, it was very difficult for linguists to convey studies on the spoken form of Dutch. Researchers had to collect their own material, which owing to money constraints was generally limited in amount.

All the recordings in the corpus were orthographically transcribed according to an extended transcription protocol (Goedertier & Goddijn, 2000; Goedertier et al., 2000) and were enriched with Part-of-Speech information. Furthermore, an automatic broad phonetic transcription together with word-based segmentations is available for all the words. For about one million words additional annotation layers containing more detailed information were generated, such as a manually verified broad phonetic transcription (Gillis, 2001; Goddijn, 2003), manually checked word segmentation (Binnenpoorte, 2002; Martens et al.,

2002) and a syntactic annotation (Van der Wouden, 2002). A small portion, about 250 K words also received a prosodic annotation (Buhmann et al., 2002; Marsi, 2003). The one million word sub corpus, the *core corpus*, is supposed to be a cross-section of the whole corpus with respect to the original design.

1.5.1 Design – recording settings and speech styles

The design of the corpus was guided by a number of considerations (Oostdijk, 2002; Boves & Oostdijk, 2003). First, the diverse group of potential users had different requirements regarding quality and quantity of the data, the number of speakers, degree of detail, etcetera. Second, there were practical constraints that influenced the ultimate design. These constraints mainly concerned the budget available for collecting and annotating recordings. The proposed corpus design is explained in terms of the various recording situations. The categorisation of the recording situations can be described by four parameters, i.e. number of speakers, private or public, degree of preparedness, and possibility of contact between speaker and listener(s). A recording situation is considered as public if the recording was explicitly meant to be heard by some audience on site or at a later period in time. The possibility of contact between speaker and listener expresses whether the listener had access to other ways of communication from the speaker besides the spoken words, such as gestures, and was possibly able to influence the behaviour of the speaker, for example by means of their postures – or perhaps also by means of verbal responses. In total fifteen different recording settings were ultimately included in the CGN. The categorisation is based on criteria that were defined a priori; the situation in which the speech is recorded determines the classification of the speech style. This is based on the assumption that each situation elicits different data; different with respect to syntactic structures, use of words, pronunciation of words, prosody, etcetera. To what extent the different recording settings affect the pronunciation is part of the exploration of spoken Dutch as described in this thesis.

In the experiments described in the following chapters, we refer to speech styles by maintaining the names and descriptions from the categorisation of the recording settings as used in the CGN project. Table 1-1 displays the seven settings we used in the experiments and the accompanying values for the different features.

Table 1-1 Recording settings used in experiments with distinctive features from CGN

	# of speakers	public (Y,N)	preparedness (-,+)	contact (Y,N)
read speech	1	Y	+	N
lecture	1	Y	+	Y
broadcast monologue	1	Y	+	N
interview	2	N	-	Y
lessons	2+	Y	-	Y
telephone conversation	2	N	-	N
face-to-face conversation	2+	N	-	Y

Some comments on this table are in order. First, the face-to-face conversations can sometimes be considered as multi-logues; more than two speakers can be involved in the conversation. The same holds for lessons where a teacher and some students discuss several subjects. Second, the interviews we selected for our experiments were not meant for broadcasting, while others in the CGN were. Third, the degree of preparedness is difficult to capture in binary values. The '+' here means 'more or less prepared'. For instance, a lecture is prepared at least for content and to a lesser extent for the actual words used during the lecture. The same holds for the broadcast monologues we selected, whereas read speech has the highest degree of preparedness. Lessons, for instance, can be considered as prepared during oral tests while during a group discussion this is much less the case. In Table 1-1 the different recording settings are ranked on the 'estimated' degree of preparedness, although it is hard to predict how this affects pronunciation in the different recordings.

1.5.2 Phonetic transcription procedure in the CGN

In the experiments described in chapters 3, 4, and 5 of this thesis, the broad phonetic transcriptions of the CGN (both the automatically generated and the hand-crafted ones) play an important role. Hence a short description of the procedure followed during the actual production of these phonetic transcriptions of specifically the Northern Dutch part is in order. Figure 1-2 is a schematic overview of this procedure.

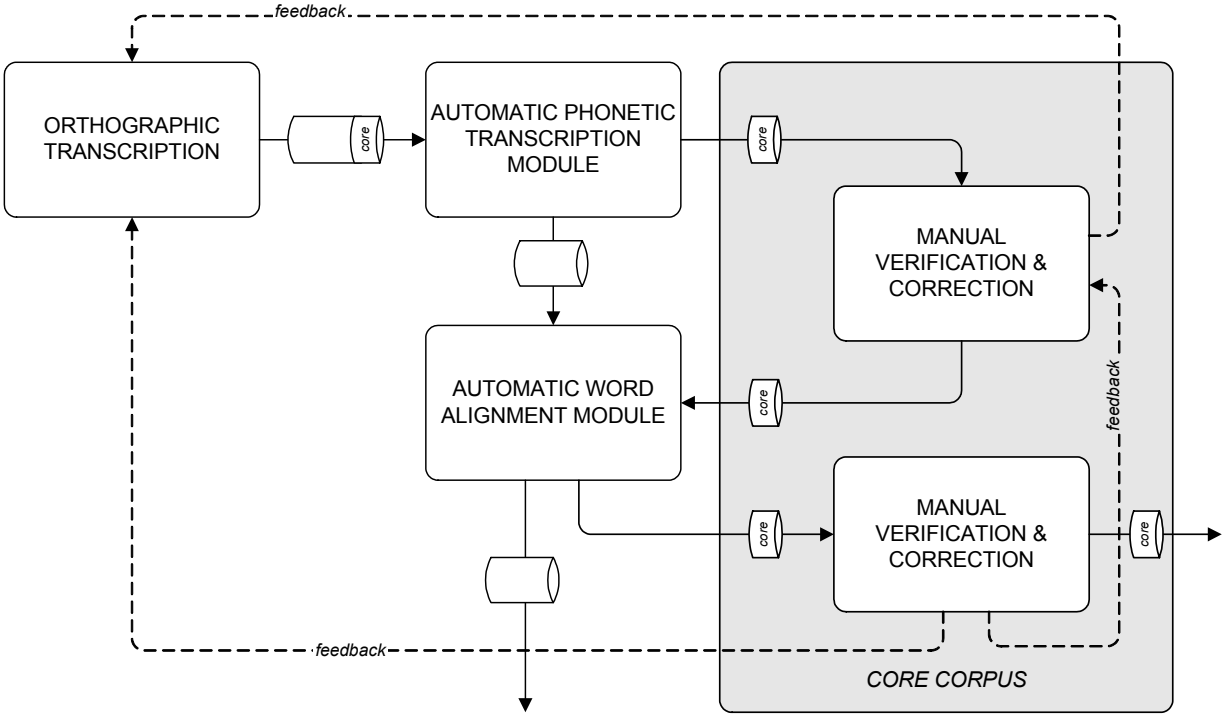

Figure 1-2 Schematic overview of phonetic transcription procedure in CGN

The orthographic transcription was input for the automatic phonetic transcription module, see for more details chapter 2 and Binnenpoorte & Cucchiarini (2003). For the part of the CGN contained in the *core corpus*, the output of this transcription module was manually verified and corrected, see chapter 4, whereas the automatic transcriptions of the rest of the data remained unverified. Both the manually checked broad phonetic transcriptions of the *core corpus* and the automatic broad phonetic transcriptions from the automatic module were input for the automatic word alignment process (Martens et al., 2002). In this module, each orthographic word was time-aligned to the speech signal based on the phonetic transcriptions, either the manually verified or the automatically generated ones. The word alignments made on the basis of the manually verified transcriptions were subsequently checked and corrected in another manual verification module. During the two manual verification phases it was possible that transcribers discovered flaws, errors, and inconsistencies in the data. A feedback procedure was designed, represented by the dashed lines in Figure 1-2, such that the bugs could be solved. Finally, the manually checked word alignment in which pauses between words were marked as well, served as input to the prosody module. The unverified automatic word alignments were released as such.

1.6 Thesis outline

In the following chapters the issues and approaches presented in sections 1.3 and 1.4 are elaborated on. First, in chapter 2 we present and test a bootstrap transcription procedure that was designed to improve automatic transcription generation and to obtain new phonological knowledge about speech styles that have not yet been investigated in depth. The proposed procedure also comprises an evaluation procedure to establish transcription quality. Subsequently, in chapter 3 the potential of pronunciation modelling based on variants observed in a transcribed corpus is examined; here, a medium-sized phonetically transcribed corpus of spontaneous speech serves as knowledge source. Using the variant-based approach, we attempt to improve the quality of automatic transcriptions of spontaneous speech. The research described in chapter 4 aims at evaluating the phonetic transcription procedure that consists of manually verifying and correcting example transcriptions, which is applied in many large speech corpora projects. For this purpose we establish the quality of human-made transcriptions by means of several quality measures. Finally, in chapter 5, we actually use a manually transcribed corpus of spontaneous speech to obtain new knowledge. We investigate whether words occurring in frequent word sequences, which are referred to as Multiword Expressions, exhibit different pronunciation patterns when they occur in such sequences as opposed to when they occur in other contexts.

In summary, chapters 2, 3, and 4 are more methodological in nature, opposed to chapter 5 in which a specific phenomenon is subject of investigation. Chapters 2 and 3 are concerned with both automatic phonetic transcription generation methods and measures to establish transcription quality. Chapter 4 focuses on manual phonetic transcription generation and evaluation methods. Chapter 5 is an exploratory study on a specific phenomenon in spontaneous speech.

A PROCEDURE FOR THE PRODUCTION OF PHONETIC TRANSCRIPTIONS OF LARGE SPEECH CORPORA

CHAPTER 2

Reformatted from:
Diana Binnenpoorte, Catia Cucchiarini and Lou Boves. A procedure to optimise the production of phonetic transcriptions of large speech corpora. Submitted to *Language Resources and Evaluation.*

The study explores the possibilities of automatic techniques for the production of phonetic transcription of large speech corpora. We argue that good quality automatic transcription can facilitate or even replace expensive and time-consuming human transcription, which are prone to subjectivity and inconsistency. In the first part of the paper, we present an iterative procedure designed to improve the efficiency of generating transcriptions of large speech corpora and at the same time to obtain new linguistic knowledge with respect to the nature and frequency of phonological processes in various speech styles.

In intermediate cycles transcription quality is measured and compared to a threshold set on the basis of inter-human agreement scores obtained in similar tasks. Once the threshold is reached, the transcription is considered to be of sufficient quality, meaning that more complicated automatic techniques can be omitted. Besides continuous quality assessments in each cycle, a detailed analysis is performed to obtain information that can subsequently be used to improve the automatic transcription where necessary. In the second part of the paper, we demonstrate the procedure with real-life data from a large speech corpus, the Spoken Dutch Corpus.

2.1 Introduction

In the past few years, many large speech corpora (LSCs) have become available for various languages, e.g. Switchboard, the SpeechDat corpora, and Verbmobil. Such large corpora are extremely valuable for developing applications and conducting linguistic research, because they contain huge amounts of 'real-life speech', which differs essentially from the 'laboratory speech' that has been used in much of the speech research carried out so far. For research and application purposes the speech in such corpora needs to be annotated at various levels, depending on the specific goal of the research or application. The annotation levels can vary from a basic orthographic transcription to a very detailed phonetic transcription or a syntactic annotation. Every type of manual annotation of speech signals is very time-consuming and costly: the more detailed the annotation, the higher the cost. In particular, phonetic transcriptions are known to be extremely labour-intensive and therefore expensive. Demuynck et al. (2002) report that one minute of semi-spontaneous speech takes an experienced transcriber about forty minutes to transcribe on a broad phonetic level. This raises the question whether manual phonetic transcriptions of LSC can always be justified. Alternatively, an automatic annotation might appear to be equally valuable, yet much less expensive. Therefore, we must address the question if and how a computer can

be deployed to obtain automatic phonetic transcriptions that are accurate enough. We will come back to the definition of "accurate enough" in section 2.2.2.

There are at least two different approaches to the problem of automatic phonetic transcription: one relying on processing large amounts of speech data attempting to optimize some formal criterion like the likelihood of the observed speech data given a hypothesis about the transcription (cf. e.g. Cremelie & Martens, 1999; Kessens et al., 2003; Schiel et al., 1998), and another based on the application of phonetic and phonological knowledge. The research reported in this paper adheres to the knowledge-based approach, but it also attempts to extend existing knowledge about phonetic and phonological processes with quantitative information that is derived from speech data. It is therefore a combination of knowledge-based and data-driven, as explained in Strik & Cucchiarini (1999, p. 231).

Much knowledge about phonetic and phonological processes is already available, and therefore can be employed in the generation of automatic transcriptions of large speech corpora (cf., Knowles, 1994). However, this knowledge is mainly based on analyses of laboratory speech (Booij, 1995; Cutler, 1998), while the LSCs that are compiled nowadays contain considerable amounts of spontaneous speech. Spontaneous speech is still under-researched, and some LSCs have been created specifically to fill this gap in linguistic research. Furthermore, much of the phonological knowledge has been obtained through introspection, a valuable method of analysis that is very suitable to investigate slow speech, or speech styles characterised by a high degree of monitoring, such as formal speech. However, this method may be less fruitful for fast speech and less monitored speech styles like casual or conversational speech, which are less amenable to introspection. Finally, the problem with many of the available phonological descriptions is that they indicate in rather general terms which processes may be expected in different types of speech, but fail to provide precise and quantitative information about the conditions under which specific processes are more or less likely to apply. However, quantitative information on the frequency of application of the various phonological processes is crucial for modelling them for automatic transcription generation (cf., Schiel, 1999). In addition, there are many languages for which very little information on phonetic and phonological processes is available, which of course makes the generation of knowledge-based automatic phonetic transcription even more difficult.

For these reasons, we are dealing with a sort of vicious circle: phonetic transcriptions are essential to conduct research on spontaneous speech phenomena, but manual transcriptions are too costly, which invokes the necessity of an automatic transcription technique that, in turn, is difficult to develop because the knowledge required is not available.

In this paper we propose a procedure intended to break the circle and to solve the dilemma outlined above. The aim of this procedure is twofold: generating knowledge-based automatic transcriptions of large speech corpora and, at the same time, gathering knowledge about the nature and frequency of application of phonological processes in real-life speech. This procedure is based on a bootstrap method of successive cycles in which the knowledge about phonological processes gathered in one cycle is implemented in the following ones. Whereas the procedure itself is intended to be generally applicable for any LSC, the precise implementation and the extracted information is, of course, data-specific and dependent on the specific language of the LSC in question. To avoid the risk of identifying peculiarities of the data used in the bootstrap procedure instead of more common speech phenomena, a validation scheme by human experts is provided during each cycle.

Thus, the aim of this paper is to show how existing techniques for knowledge-based automatic phonetic transcription can be usefully combined in a procedure for automatic transcription generation and knowledge extraction that can subsequently be applied to obtain high quality broad phonetic transcriptions of LSCs. To demonstrate the feasibility of this procedure for LSC exploration in general, we will test it on an existing corpus of spoken Dutch.

In this paper we first address the way in which transcription quality can be assessed, and the issue of how to determine whether a transcription is 'good enough'. We will then proceed to explaining the bootstrap procedure that we propose in detail. Finally, we report on experiments that illustrate how our procedure can be put into practice. Although our approach is intended for LSC exploration, for the experiments we use fragments extracted from a completely annotated LSC, because this allows us to check the working of our procedure in detail.

2.2 Assessing transcription quality

2.2.1 Basic notions: reference transcription, transcription quality

Since LSCs usually contain an orthographic transcription of the speech material and a pronunciation lexicon with canonical transcriptions, a rudimentary form of automatic phonetic transcription can be obtained by simply concatenating the canonical representations of the orthographic words. Although such a form of automatic phonetic transcription may suffice for some applications such as training an automatic speech recogniser, it is obvious that it will not be extremely accurate. In particular, concatenations

of canonical forms will not accurately reflect spontaneous speech, as this speech style is known to deviate considerably from canonical forms (Engstrand, 1992; Kohler, 1998; Swerts et al., 2003).

To determine how precisely an automatic transcription represents the actual speech signal, we need a reference to evaluate the automatic transcription. In phonetic research, the difficulties of obtaining such a reference 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 (Cucchiarini, 1993). When making phonetic transcriptions, human transcribers are susceptible to bias by their own hypotheses and expectations. In addition, they are likely to make mistakes owing to fatigue and/or loss of concentration. Consequently, human-made phonetic transcriptions may contain an element of subjectivity. In an attempt to reduce human subjectivity phoneticians have been looking for procedures that can approach a true reference transcription. In Shriberg et al. (1984), a consensus transcription is suggested as a possible alternative. A consensus transcription is made by two or more experienced transcribers after they have agreed on each individual symbol. Other transcriptions can then be evaluated by comparing them to the consensus transcription of the same material on a symbol-by-symbol basis. A dynamic programming algorithm can be used for this purpose. The extent of deviation can be used to measure the quality of a given transcription; the quality is expressed using the level of (dis)agreement.

2.2.2 When is an automatic transcription good enough?

An automatic transcription as is proposed in this paper can only replace a human-made transcription if its quality is good enough. In the suggested procedure, a consensus transcription will serve as the reference transcription with which the automatic transcription is compared to determine the level of (dis)agreement. This level of (dis)agreement indicates the quality of the automatic transcription in question, but in order to decide whether the quality is "good enough", the (dis)agreement level must be compared to some sort of criterion or threshold. In some cases the threshold can be derived from external requirements. This is the case if the goal for making the transcriptions is well defined, if the contribution of transcription quality to achieving the goals is well understood, and if the degree to which the goal has been achieved can be established with independent measurements. However, LSCs are typically intended to serve multiple purposes. Therefore, it is impossible to define a minimum quality level on the basis of specific purposes. In the absence of an externally defined threshold, and since an automatic transcription is intended to replace a human-made transcription, the performance of a

single expert transcriber is the best quality one can hope to achieve. It follows that the degree of (dis)agreement between two human experts working independently and from scratch is a good indicator of the best possible transcription quality we can get from human labour. If automatic transcription can provide the same (or perhaps better) agreement with human made transcriptions, we consider this as 'good enough' for a multipurpose LSC Table 2-1 gives an overview of agreement scores for a number of different speech styles reported over the last decade. In interpreting these scores, it should be taken into account that the values are not based on comparing the transcriptions of phoneticians who worked *independently*. This is because the transcribers started from an example transcription in all experiments summarised in Table 2-1.

Table 2-1 Overview of inter-labeller agreement for various speech styles reported in various studies.

reported in	agreement percentage	speech material	task of labeller	# labellers
Eisen, 1993	varied between 70% (glottal stops) - 96% (fricatives)	read speech (Phondat)	SAMPA broad transcription	4
Greenberg, 1998	between 75% - 80%	spontaneous speech (Switchboard)	annotation with Arpabet extended with diacritics	8
Kipp et al., 1996	between 93.1% - 94.4%	read speech (Phondat II)	SAMPA broad transcription	at least 3
Kipp et al., 1997	between 78.8% - 82.6%	spontaneous speech (Verbmobil)	SAMPA broad transcription	3
Raymond et al., 2002	between 73% - 76%	spontaneous speech (interviews in Buckeye corpus)	verify given transcription in DARPA phonetic alphabet	4
Wester et al., 2001	between 75% - 87%	extemporaneous speech (VIOS)	SAMPA broad transcription; choosing most probable variant	9

In Binnenpoorte et al. (submitted) it is shown that if human transcribers are asked to edit an example transcription, the agreement levels may be artificially high. This is due to the influence the example transcription can have on the judgements of the human transcribers. In fact, transcribers who leave the given example transcription intact are more likely to produce a 100% level of agreement than critical transcribers. Consequently, if transcribers had to start from scratch, without some default transcription at their disposal, substantially

lower agreement scores might have been obtained. Therefore, the agreement measures summarized in Table 2-1 are most probably biased towards the high end of the scale.

As the table shows, the highest inter-transcriber agreement score reported for read speech is 94.4%, and for spontaneous speech 82.6%. Wester et al. (2001) report a maximum agreement score of 87%, but the speech material that was transcribed can be considered as less conversational than the speech analysed in the other studies on spontaneous speech in Table 2-1.

In actual practice, obtaining a reliable estimate of inter-transcriber agreement is difficult and time-consuming, if only because reliable estimates would require the deployment of a large number of transcribers. Moreover, the level of agreement is dependent on, for instance, the languages, the quality of the speech signal, the type of phoneme set, and the level of experience of the transcribers. For these reasons, we propose to estimate transcription quality from the agreement between a single (automatic) transcriber and a consensus transcription. Extrapolating the figures in Table 2-1 to this condition, it seems reasonable to consider 90% agreement between a consensus transcription and an automatic transcription for read speech as the threshold for "good enough". For conversational speech, we set the threshold at 80% agreement. For speech styles that are in between read speech and conversational speech with respect to expected spontaneous speech effects, we will require agreement between automatic transcriptions and a consensus transcription in the range between 80% and 90%.

In addition the suspicion that the values in Table 2-1 are biased to the high end of the scale justifies slightly lower criterion values. Moreover, automatic transcriptions have a substantial cost advantage over manual transcriptions, are more consistent, and certainly more reproducible than the transcriptions of individual experts.

2.3 Design of the bootstrap procedure

Once a criterion for deciding whether an automatic transcription is good enough has been set, it can be determined whether an initial automatic transcription can replace human-made transcriptions. If this is not the case, measures must be taken to improve the as yet defective automatic transcription such that it becomes a more accurate representation of the speech that was actually realised. In our research, we intend to use existing and develop new explicit phonetic knowledge to improve the automatic transcription. In doing so, we are confronted by the fact that most of the existing phonetic knowledge is derived from read speech. To extract new information on spontaneous speech, we compare automatic transcriptions based on existing phonetic knowledge with consensus transcriptions of a sufficiently large corpus of spontaneous speech. We will consider systematic discrepancies

as ‘spontaneous speech rules’, which can subsequently be employed to generate new, and hopefully improved automatic transcriptions.

For this study the consensus transcription was made from scratch by two experienced phoneticians. Since a consensus transcription is very time-consuming and therefore expensive, only small samples of an LSC can be transcribed in consensus mode. This obviously limits the extent to which spontaneous speech rules can be derived by comparing rule-based predictions with actual transcriptions. We will return to this issue in the discussion (section 2.6). To avoid extraction of phonological knowledge that only applies to a small sample of utterances, two medium-sized, representative samples were created. The first set, referred to as the development set (D-set), is used to discover processes and to formulate rules. The validation set (V-set) is then used to establish whether the newly detected rules generalize beyond the corpus from which they were derived.

The procedure we propose is a step-by-step bootstrap method that should not only yield better transcriptions after each step, but also more knowledge about spontaneous speech effects that can be expressed as deterministic and/or probabilistic rules. Each step incorporates a validation phase to measure the quality of the automatic transcription. Subsequently, a detailed analysis reveals which phonological processes underlie the discrepancies between the automatic transcription and the reference transcription. In the following step, measures are taken to reduce these discrepancies and the results are then evaluated in a new quality measurement. An advantage of this step-wise method is that after each validation it is possible to determine whether the transcription obtained at that point is already of sufficient quality for certain speech styles, so that additional effort is not required.

2.3.1 The cycles and stages of the bootstrap procedure

The procedure we suggest is presented schematically in Figure 2-1. It consists of four successive cycles. Each cycle is composed of four stages:

- stage 1: *Generation*: an automatic transcription is generated;
- stage 2: *Validation*: assessment of automatic transcription to determine whether the transcription quality is already good enough;
- stage 3: *Diagnosis*: identifying the nature of the discrepancies between automatic transcription and reference transcription;
- stage 4: *Remedy and decision*: formulation of rules to be implemented in the succeeding generation stage to improve the quality of the automatic

transcription. Depending on the nature of the newly obtained knowledge, a decision is made about the manner of implementation, i.e., which cycle to enter. Only the first cycle cannot be re-entered.

The precise implementation of the stages is different in the four cycles, as will be explained below.

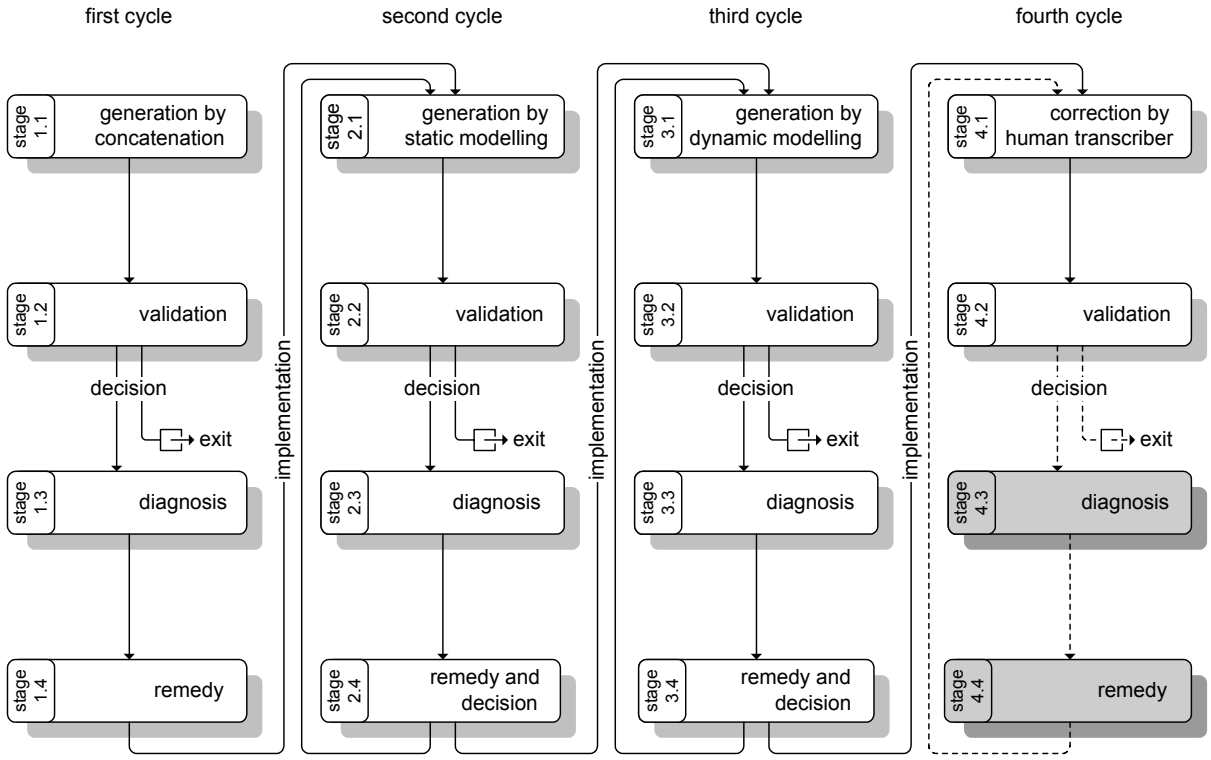

Figure 2-1 The transcription procedure with the various stages. The grey coloured stages are normally skipped.

Generation

Each cycle begins with a generation stage: a transcription is generated, either automatically (cycles one to three) or manually (cycle four). The method of transcription generation differs per cycle. Complexity and effort needed to develop the generation procedures increase from cycle one to cycle four.

In the first cycle, a simple lexicon look-up procedure is applied, *generation by concatenation*, the cheapest and easiest way to produce automatic phonetic transcriptions, provided that an orthographic transcription and a lexicon with canonical pronunciations are available. This first step serves as a starting point from where the transcriptions can be improved.

In the second cycle, the generation is implemented through, what we will call, *static modelling* of pronunciation variation. Static modelling means that the most frequent variant is chosen instead of the canonical form for a specific phonological process, irrespective of the actual speech signal. Static modelling can also be considered as a grapheme-to-phoneme conversion procedure that can be optimized for a range of speech styles. In terms of expected agreement with a consensus transcription that is based on the actual speech signals, such a deterministic method only makes sense for processes that are applied either very frequently, i.e., in more than 75% of the cases, or very infrequently, in less than 25% of the cases.

In the third cycle, generation occurs through *dynamic modelling* of pronunciation variation. Dynamic modelling is a probabilistic method of transcription in which an automatic speech recognizer (ASR) is used to choose the transcription that best matches the speech signal given a list of possible pronunciation variants. In principle, dynamic modelling can be applied to all phonological processes, but we prefer to start with static modelling, since this helps to discover deterministic knowledge that can easily be implemented without using an ASR. We limit dynamic modelling to the problems that defy attempts to apply deterministic procedures.

Finally, the fourth cycle is called *human correction*. This is the most time-consuming and extensive stage of the whole procedure, and the idea behind our approach is to try to avoid this stage as much as possible, by maximally exploiting automatic techniques. As a matter of fact, in our procedure, this stage is entered when rule-based improvement of automatic transcription is no longer possible.

Validation

During the second stage of each cycle, the validation stage, the quality of the automatic transcription is measured in order to determine whether the quality is good enough. The automatic transcription is evaluated automatically by aligning it with the reference transcription of the same speech material, the V-set. The latter is a consensus transcription made by expert transcribers. Transcription quality is expressed as percentage of disagreement on symbol level, which is one hundred percent minus the agreement percentage (see section 2.4.3). The criterion values for read speech and spontaneous conversations were set at 10% and 20% disagreement, respectively.

Diagnosis

In the diagnosis stage, the automatic transcription is examined to determine in what respects it deviates from the reference transcription. This evaluation is required to formulate rules for possible improvements in the following stage, the remedy stage. The V-

set cannot be used for this purpose, because one would run the risk of modelling training-data-specific processes. Therefore, a second set is required, the D-set, for which a consensus transcription is also required.

In the diagnosis stage during the second cycle, we search for speech style dependent phonological processes that might be used to improve the quality of grapheme-to-phoneme conversion, whereas in the third cycle the search is for rules that can be used to generate plausible pronunciation variants.

Remedy and decision

In the remedy and decision stage rules are actually formulated for improving the automatic transcription, and a choice is made on how to implement these rules. The iterative bootstrap procedure depicted in Figure 2-1 is designed in such a way that the second and third cycles can be repeated. Thus, if at the end of the first completion of the second cycle it appears that there are systematic grapheme-to-phoneme relations that are not yet covered by the rules, these phenomena are added to the rules, perhaps only for a specific speech style. The cycle is then repeated. If no additional rules seem to exist, the procedure moves forward to the third cycle, most likely with an initial set of rules for generating plausible pronunciation variants. The third cycle can be repeated, each time with more speech style specific rules for generating additional variants, until no new rules can be discovered. Then, the procedure advances to the fourth and last cycle.

2.4 The design of the experiment

In this section we describe the design of the experiment in which the procedure sketched in section 2.3 is applied up to cycle four. The aim of the experiment is to test whether the cyclical procedure is an efficient way to generate automatic phonetic transcriptions for LSCs, while obtaining additional knowledge about the phonological processes that characterise the various speech styles.

2.4.1 Speech material

The real-life speech database to which we applied our procedure is the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN), a large corpus (about 9 million words) of Dutch as spoken in the Netherlands and Flanders, containing speech from a great variety of socio-situational settings (Oostdijk, 2002). Since the CGN was completed in 2004, we did not apply our procedure to transcribe it, but to illustrate and evaluate our approach. The

two sample sets, the V-set (validation set) and the D-set (development set), were selected from the Northern Dutch part of the corpus.

Validation set: V-set

The sample sets must be representative of the whole corpus to be transcribed. The CGN is designed to represent speakers of different age, gender, and region of origin, recorded in various situations, which resulted in speech styles ranging from formal to conversational. When we defined the samples, we ensured that they would cover this variation in speaker characteristics as well as the different speech styles in the CGN. The V-set consisted of speech material from 27 different speakers, 13 male and 14 female. The speakers were 21 to 60 years of age at recording time and came from various regions in the Netherlands.

The speech material was collected from 16 different fragments from the CGN extracted from the following components: read speech (RS), broadcast monologues (BM), spontaneous telephone conversations (ST), and spontaneous face-to-face conversations (SC). Thus, the fragments sample the most formal speech style in the corpus (RS), the most informal style (ST and SC) and a style between the two extremes (BM). Most likely, BM is closer to RS than to ST and SC.

In the RS fragments, trained speakers read novels aloud in a studio environment. The BM fragments were also produced by speakers used to speaking in public. In contrast, no professional speakers were involved in the SC and ST fragments. The SC fragments were recorded in a home environment where at least two speakers were having a conversation while engaged in daily activities, such as eating, playing a game, or watching television. The speakers in the ST fragments were friends and relatives; during the conversations the speakers were not engaged in activities other than the phone call (Oostdijk, 2002). The most relevant statistics of the V-set are presented in Table 2-2. The counts for number of utterances, words, and phonemes are based on the consensus transcription.

Table 2-2 Statistics of the speech material in the V-set: number of utterances, words, phonemes, duration in seconds, number of female speakers (F), and number of male speakers (M).

speech style	# utterances	# words	# phonemes	duration (s)	#F	#M
RS	74	416	1537	125.06	2	1
BM	79	519	2076	156.01	2	2
ST	175	952	2868	251.54	5	5
SC	158	696	2253	210.14	5	5
total	486	2583	8734	742.75	14	13

Development set: D-set

The D-set also consisted of 16 different fragments, representing four components in the CGN: read speech (RS), lectures (LC), interviews (IN), and spontaneous conversations (SC). Again, the fragments represent the extremes on a formal - informal scale (RS and SC), and two styles that are in between the extremes (LC and IN). IN and SC are dialogues, thus containing speech from at least two different speakers. The lectures were recorded during the actual lectures, in various environments, with speakers used to talking in public. The speakers in the interviews were teachers of Dutch who were talking to an amateur interviewer. The recordings were not made in the classroom, but in acoustically treated studios. The speech material in the D-set was produced by twenty different subjects, nine female and eleven male speakers, who originated from different regions in the Netherlands. There was no overlap between the speakers of the D-set and the V-set. Table 2-3 gives an overview of the quantitative characteristics of the D-set.

Table 2-3 Statistics of the speech material in the D-set: number of utterances, words, phonemes, duration in seconds, number of female speakers (F), and number of male speakers (M).

speech style	# utterances	# words	# phonemes	duration (s)	#F	#M
RS	140	673	2749	223.51	3	2
LC	101	704	2638	248.56	0	5
IN	72	428	1513	141.36	2	2
SC	79	465	1401	124.46	4	2
total	392	2270	8301	737.89	9	11

In order to avoid the risk of finding spurious regularities, care was taken to make the D-set and V-set as different as possible, without making it impossible to compare and generalize between the sets. Since one of the aims of our procedure is to obtain knowledge on speech processes that can be used to model these processes with a view to obtaining better automatic transcriptions of speech, it is important to know whether the knowledge extracted generalizes to other exemplars of the same speech style. For this reason, we expressly decided to choose different components of the CGN as representatives of speech that is in between well-prepared (RS) and unprepared (ST and SC): in the D-set we included lectures (LC) and interviews (IN), and in the V-set we included broadcast monologues (BM). In addition, to avoid the confusion of speaker specific phenomena with general phonological processes, the speech fragments were selected such that no speaker occurred in both sets. As a consequence the topics discussed in the fragments also differed.

2.4.2 Consensus transcriptions

Two experienced phonetically trained transcribers made a consensus transcription of all speech in the V-set and D-set. They listened to the speech material and transcribed from scratch. First, they made the transcription while working on their own, then they compared the transcriptions with each other and, finally, they reached consensus on every symbol included in the final transcript. The symbol set used was an adaptation of SAMPA for Dutch (Wells, 2004). The same set was used for the compilation of the CGN lexicon (Oostdijk, 2004a). The total of the three steps - individual transcription, comparison, and negotiation agreement - took about sixty minutes per minute of speech on average.

2.4.3 The *Align* program

To determine the distance between the consensus and the automatic transcriptions, the symbol sequences must be aligned in such a manner that the total symbol-to-symbol difference is minimal. This was accomplished with the program *Align* (Cucchiarini, 1996). *Align* employs dynamic programming techniques to find the optimal alignment between two strings. Costs for deletions, insertions, and substitutions of the phonetic symbols are determined on the basis of articulatory features such as voicing, lip rounding, and length. For example, substituting a /t/ for a /d/ (which differ only in the feature voicing) has a lower cost than substituting a /t/ for an /x/ (which differ in manner and place of articulation), see Appendix A1 and A2. *Align* outputs the number of substitutions, deletions and insertions on symbol level and expresses the results in percentages of disagreement:

$$\frac{S + D + I}{\#symbols} * 100\%$$

where S is the number of substitutions, D the number of deletions, and I the number of insertions. The denominator, the number of symbols, is the number of symbols in the reference transcription. In addition to an overall disagreement measure, *Align* also provides information on the nature of the differences, i.e. differences in articulatory features, between two transcriptions which is used to develop remedies.

2.5 The cycles

2.5.1 First cycle

The first cycle in the procedure that we propose boils down to analyzing the extent to which general (speech style independent) grapheme-to-phoneme conversion predicts actual pronunciation. In our experiment grapheme-to-phoneme conversion was implemented in the form of a straightforward lexicon look-up.

Generation - Stage 1.1: Concatenation

The CGN has a lexicon that comprises all the words in the corpus (Oostdijk, 2004b). This lexicon contains the orthographic and the corresponding canonical phonetic representation of each word, obtained using TREE-TALK (Daelemans & Van den Bosch, 2001; Hoste et al., 2004), a grapheme-to-phoneme converter trained on CELEX (Baayen et al., 1995) by means of memory-based learning. In the phonetic representations in CELEX, all obligatory word-internal phonological processes were applied, but word-internal processes that are considered as optional in the present state of the knowledge were not. The transcriptions resulting from the first stage of the first cycle are referred to as PT1; they are generated for both V-set and D-set.

Validation – Stage 1.2

The automatic transcription, PT1, of the V-set is compared with the reference transcription of the V-set using the *Align* program. In Table 2-4, the frequency of the substitutions, deletions, and insertions in the V-set is displayed for the four speech styles. To clarify, a deletion means that a symbol was present in the reference, but not in the automatic transcription, and vice versa, an insertion means that a symbol was present in the automatic transcription and not in the reference transcription.

Table 2-4 Percentages disagreement of PT1 after validations for all speech styles (V-set).

speech style	substitutions	deletions	insertions	total
RS	8.3	0.5	5.5	14.3
BM	8.6	1.6	6.0	16.2
ST	10.7	1.5	11.9	24.1
SC	11.1	0.9	14.3	26.3

The total percentages of disagreement displayed in the rightmost column indicate that for all parts of the V-set the quality of PT1 is below the threshold defined in section 2.2.2. Therefore, it is necessary to investigate whether and how PT1 can be improved. To this end, we used the D-set to analyse the discrepancies and define a remedy.

Diagnosis – Stage 1.3

In the diagnostic stage, the automatic transcription is compared with the reference transcription of the D-set. Again, we used the program *Align* to obtain our first quantitative results. In Table 2-5, the percentages of substitutions, deletions, and insertions are displayed.

Table 2-5 Percentages disagreement of PT1 after diagnosis for all speech styles (D-set).

speech style	substitutions	deletions	insertions	total
RS	7.1	0.9	4.2	12.2
LC	8.9	0.4	9.2	18.5
IN	7.6	1.1	11.4	20.1
SC	10.8	0.9	13.1	24.8

V-set and D-set differ in the speech components and their size, measured using the number of words (and phonemes and utterances). RS and SC, however, are present in both sets, but show different disagreement percentages. The SC fragments in the two sets do not only vary in size, but more importantly, they sample fragments from different speakers. Given the relatively small size of the sets, the differences in total percentage disagreement do not come as a surprise. However, the disagreement percentages for the RS and SC fragments from the two sets do not differ significantly ($p \geq .05$). Besides, the relative contribution of deletion, insertion, and substitution errors is highly similar in the V-set and D-set for these two speech styles.

In addition to computing insertions, deletions, and substitutions of phones, we also analysed the nature of the discrepancies between PT1 and the reference transcription of the D-set. The error types were divided into two main categories: a) errors at word boundaries, and b) word-internal errors, to distinguish between cross-word and word-internal processes. A more detailed analysis of the errors is presented in the next section.

Remedy and decision – Stage 1.4

For all the speech styles, a list of the most frequent errors in both word-boundary and word-internal position was compiled. As the D-set is rather small, it is difficult to define a strict frequency threshold below which a phenomenon is no longer of interest. Therefore, the errors with the highest overall frequency within a speech style were selected for further analysis. To avoid including speaker-specific phenomena, only those errors observed for more than one speaker were considered. Table 2-6 lists these errors accumulated per speaking style. Their relative frequency is calculated to determine whether to address them through static modelling or through dynamic modelling.

Table 2-6 Listing of highly frequent, non-speaker specific errors (in Dutch SAMPA notation).

	substitutions		deletions		insertions	
	boundary	internal	boundary	internal	boundary	internal
RS	f-v	x-G	h		n	r
	s-z	E-@			t	
	t-d				r	
LC	f-v	x-G			n	@
	s-z	@-A			r	r
	z-s	@-E			t	l
	t-d					
	d-t					
	@-I					
	@-E					
IN	f-v	x-G			@	r
	t-d	@-A			n	@
					r	
					t	
SC	f-v	x-G			@	r
	s-z	@-A			n	@
	z-s				t	
	@-E				r	
	t-d					

Table 2-6 shows more frequent errors at word boundaries than within words, especially in lectures and spontaneous conversations. Substitution and insertion errors are clearly much more frequent than deletions. Finally, voiced-voiceless confusions (and vice versa) are responsible for a large proportion of the substitutions. Since there are more cross-word voice substitutions than within-word voice substitutions, we first focussed on analysing cross-word errors. The large number of cross-word voice substitutions can be accounted for by a well-known process, cross-word voice assimilation (Booij, 1995).

Table 2-7 shows the frequency (column 2 gives the absolute frequencies and column 3 the percentages of all substitution errors) of word-boundary voice substitutions per speech type. The F*rel* column indicates the relative frequency of the cross-word voice assimilation processes. F*rel* 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.

Table 2-7 Word-boundary voice substitutions and relative frequency of voice assimilation.

speech style	voice substitutions		F*rel*
	#	%	%
RS	82	83.7	88.7
LC	95	71.9	86.7
IN	46	66.7	94.7
SC	60	63.8	92.9

The data in Table 2-7 indicate that word-boundary voice substitutions are relatively frequent and that this process (cross-word voice assimilation) is frequently applied in all four speech styles. The high values of F*rel* suggest that if the variant with the voice-assimilated phoneme were chosen, PT1 would resemble the reference transcription in the D-set more closely.

In Table 2-6, more insertion processes are found at word boundaries than word-internally. This can partly be attributed to cross-word degemination processes that are very common, but were ignored in PT1. In contrast, word-internal degemination was already applied in the lexicon.

Although most of the highly frequent errors occur at word boundaries, there is one word-internal substitution in Table 2-6 that was also taken into consideration, namely the substitution of /x/ (voiceless velar fricative) for /G/ (voiced velar fricative), since the F*rel* of this process varies between 84.8% and 79.1%. These high figures indicate that choosing the most frequent variant will improve PT1. Loss of the distinction between the voiced and

voiceless velar fricative is supported by data reported in the literature. Gussenhoven (1992), Smits et al.(2003) and Van de Velde (1996) indicate that most Dutch speakers neutralise the distinction in favour of the voiceless variant.

The following rewrite rules were formulated based on the observations in the D-set and were used in the second cycle to generate the new variants:

Progressive voice assimilation: [+fric, +voice] → [-voice] / [-son, -voice] _
Regressive voice assimilation: [+plos, -voice] → [+voice] / _ [+plos, +voice]
Degemination rule: C1 → ∅ / _ C1, where C1 is any consonant.
Devoicing the velar fricative in every context: /G/ → /x/ / c _ c, where c is any context.

We have calculated the F*rel* for all the other types of errors listed in Table 2-6. Since none of these errors had an F*rel* close to 75%, they could not be resolved by deterministic rewrite procedures and thus probably need a different approach.

2.5.2 Second cycle

Generation – Stage 2.1: Static modelling

The rules developed in the first cycle were applied to PT1 by means of rewrite rules on the concatenated canonical representations of the words; the resulting phonetic transcription is referred to as PT2.

Validation – Stage 2.2

In stage 2.1, both the V-set and the D-set were automatically transcribed into PT2. The V-set again was used to measure the quality of the transcription. In Table 2-8, the figures represent the percentages of error, substitutions, deletions, and insertions on symbol level for PT2.

Table 2-8 Percentages disagreement of PT2 after validation for al speech styles (V-set).

speech style	substitutions	deletions	insertions	total
RS	6.6	0.6	4.7	11.9
BM	7.7	1.9	5.4	15.0
ST	9.1	1.5	10.9	21.5
SC	10.8	1.0	13.4	25.2

Comparing Table 2-8 and Table 2-4 reveals that PT2 outperforms PT1 for all four speech styles. The relative improvement varies between 4.1% and 16.8%, with RS and ST showing the greatest improvement. The percentages of both substitutions and insertions have decreased for all speech styles. As predicted, static modelling of highly frequent phonological processes gives a higher level of transcription quality.

In section 2.2.2 we defined the threshold for deciding whether a transcription is good enough as 90% agreement for read speech and as 80% agreement for spontaneous speech. None of the speech styles has reached the criterion values yet, although read speech and telephone conversations actually come very close. Therefore, we continued analysing the remaining discrepancies for all four speech styles and tried to improve the quality of the automatic transcription.

Diagnosis – Stage 2.3

We aligned PT2 of the D-set with the reference transcription. The total percentages of deviation show an overall improvement compared with the results of PT1 in the first cycle, as can be seen by comparing Table 2-5 and Table 2-9.

Table 2-9 Percentages disagreement of PT2 after diagnosis for all speech styles (D-set).

speech style	substitutions	deletions	insertions	total
RS	5.1	1.1	3.7	9.9
LC	7.2	0.7	8.3	16.2
IN	6.5	1.3	10.7	18.5
SC	9.2	1.3	12.1	22.6

It is interesting to note that the total percentage of errors for RS in the D-set is lower than the criterion value of 10%. We found a decrease in both word-internal and word-boundary substitutions, as well as word-boundary insertions. However, at the same time the number of deletions at word boundaries increased slightly. This can be explained by the fact that the average F*rel* of voice assimilation at word boundaries is about 88%; thus, we have hypothesized voice assimilation in a number of cases were it did not actually apply. Furthermore, when assimilation is applied, degemination can possibly occur. In our implementation, degemination was applied in 100% of the cases in which it could have been applied; when assimilation was not justified, the degemination that followed was not correct either.

Further analysis of the discrepancies between the automatic and consensus transcriptions did not bring to light phonological processes that apply in more than 75% of

the cases where they are licensed. Thus, we have not discovered phenomena that are not attested in laboratory speech, but still appear to be important in more spontaneous speech styles. It remains to be seen whether this is due to the fact that there is already substantial knowledge about the phonetics and phonology of Dutch, or because after all the differences between read and spontaneous speech are less systematic than one might perhaps expect. In any case, our data confirm that every grapheme-to-phoneme conversion procedure for Dutch should include a number of cross-word phonological phenomena, a finding that most likely generalizes to many other languages.

Remedy and decision – stage 2.4

The discrepancies listed in Table 2-6 which are not sufficiently frequent to be modelled statically and which therefore persisted in PT2 must be resolved by applying dynamic procedures that involve the speech signal. The phonological processes underlying these discrepancies have an F*rel* approximately between 25% to 75%. These F*rel* values suggest that deterministic modelling will yield similar numbers of improvements and deteriorations. For these cases we will use rules to generate plausible pronunciation variants, and an ASR that selects the variant that fits the speech signal best. The rules, and especially the relative frequency with which the rules are applied may vary between speech styles.

The first phonological processes to address are word-final deletions of /n/, /r/, and /t/, which account for a substantial number of word-final insertions. These processes are well known in Dutch (Booij, 1995; Van den Heuvel & Cucchiarini, 2001); however, their F*rel* varies considerably over the various speech styles, especially for /t/-deletion and /r/-deletion. In RS, /t/-deletion is very rare and therefore this process is better not modelled at all for this speech style. The average F*rel* value for the other speech styles is 29.3%, which justifies a rule for generating pronunciation variants. /r/-deletion is better not modelled in RS either, because it has an F*rel* of only 9.6%. The average F*rel* for the other speech styles is 51.7%, which justifies the generation of /r/-less variants. Word-final /r/-deletion after /@/ is a special case of /r/-deletion and has an F*rel* of 57.2%, in all speech styles. Therefore, /r/-deletion after /@/ is also modelled for RS. Word-final /n/-deletion has an F*rel* value of 39.8% and is modelled for all speech styles. The following rewrite rule was applied to the words in PT2, for the speech styles to which it should have been applied:

$$\{/n/,/r/,/t/\} \rightarrow \varnothing \; / \; A \; B \; \_ \; |$$

where A B are at least two phonemes, and | represents a word boundary. The constraint that the deletion rule only applies to words with a canonical representation that contains at least

three symbols was added to avoid deletions that would have resulted in a large number of mono-phonemic forms.

Additional phonological processes to be modelled are word-internal /r/-deletion and /@/-deletion. From the literature (Van den Heuvel & Cucchiarini, 2001), we know that word-internal /r/-deletion occurs most likely in: a) a postvocalic position where the vowel is unstressed, e.g., /pAr-'ke-r@n/ → /pA-'ke-r@n/ ('*to park*'), or b) a post-schwa position, e.g., /'A-l@r-hAn-d@ / → /'A-l@-hAn-d@/ ('*all kinds of*'). However, we found hardly any word-internal /r/-deletions in RS. In the more spontaneous speech styles, /r/ is deleted in 41.6% of the times in post-schwa position, and in 35.4% of the times in other post-vocalic positions. Therefore, /r/-less variants were not generated for RS.

According to Booij (1995), /@/-deletion can be applied to the first /@/ in a word that has two consecutive syllables with /@/ in syllable-final position. The deleted /@/ must follow an obstruent and precede a liquid. E.g., /x@-'mA-k@-l@k/ → /x@-'mA-kl@k/ ('*easy*'). This process does not seem to have been applied in RS (see Table 2-6). We observed /@/-deletions in unstressed syllables, especially in the context of the prefix 'ge-' /x@/, where the /@/ was deleted in 44.0% of the cases, at least in the more spontaneous speech styles. The following rewrite rules were applied to the words contained in PT2 and stored in the recognition lexicon:

/r/ → ∅ / {vowel} _ , where the vowel is unstressed or a schwa.
/@/ → ∅ / {obstr} _ {liquid} /@/

The last phonological process to be modelled is vowel reduction, which - as Table 2-6 shows - is frequent in all speech styles except for RS. We found that vowel reduction was mainly applied to monosyllabic function words in IN and SC, and to unstressed syllables in multi-syllabic words in LC. Booij (1995) states that vowel reduction rarely applies to high vowels (/i/, /y/) or diphthongs in syllable-initial position, or in word-final syllables. Vowel reduction is preferred in open syllables in inter-stress positions, e.g. /'e-ko-no-'mi/ → /'e-k@-n@-'mi/ ('*economy*'), where vowel reduction is possible in two positions. Before the following rewrite rule could be applied, both syllable and stress information had to be added to the lexical entries:

{/a/, /e/, /u/, /o/, /A/, /E/, /I/, /O/, /Y/} → /@/ / | stress-syl | C _ | stress-syl |

where | is a syllable boundary and C any consonant. This rule is also applicable to some monosyllabic function words such as 'en' ('*and*'), 'dat' ('*that*'), 'van' ('*of*'), and 'ik' ('*I*'). We found many vowel reductions in these words, especially in the IN and SC fragments. Therefore, pronunciation variants with a schwa of these short words were added to the lexica for the ST and SC fragments.

2.5.3 Third cycle

Generation – Stage 3.1: Dynamic modelling

The utterances in the V-set and D-set were transcribed automatically using a dynamic modelling technique. An ASR in forced recognition mode (Kessens et al., 1999) established which pronunciation variant of a word best matched the speech signal. The ASR (as described in Strik et al., 1996) uses acoustic models, word-based language models and a multiple pronunciation lexicon that was enriched with pronunciation variants generated by means of the conditional rewrite rules described in stage 2.4. To ensure that the ASR did not attempt to recognise words which were not uttered at all, each utterance had its own language model and a lexicon with all relevant pronunciation variants. Because the D-set is too small to obtain reliable estimates of the relative frequency of the newly generated pronunciation variants, all variants were given the same prior probability. The resulting transcription will be referred to as PT3.

Validation - Stage 3.2

PT3 was validated by comparing it with the reference transcription, again by using the *Align* program. In Table 2-10 the results of the validation are presented.

Table 2-10 Percentages disagreement of PT3 after validation for all speech styles (V-set).

speech style	substitutions	deletions	insertions	total
RS	6.9	3.1	1.7	11.7
BM	7.0	5.5	2.0	14.5
ST	9.7	3.7	5.4	18.8
SC	11.0	4.8	8.1	23.9

By comparing the figures of Table 2-10 with the validation results of PT1 in Table 2-4 and PT2 in Table 2-8, an overall decrease in percentages of disagreement for all four speech styles is observed. The least improvement - in both absolute and relative numbers - is measured for RS. The graphs in Figure 2-2 display an overview of the transcription quality obtained for the various PTs for all the speech styles. The improvements (or deteriorations) of the PTs after each iteration are visible in the bars representing the total percentage disagreement. The improvement in transcription quality for ST is significant ($p \leq .01$) in each cycle. No significant improvements could be observed in the results of the other speech styles.

As for the types of errors, an overall decrease in the number of insertion errors can be observed. The phonological processes we implemented for PT3 mainly concerned deletions of certain phonemes in specific contexts. The implementation did indeed lead to fewer insertions for all speech styles. On the other hand, the number of deletion errors increased for all speech styles. This means that the ASR prefers shorter pronunciation variants to longer ones. The number of substitutions increased marginally in some speech styles.

Considering the overall results and the quality threshold defined in section 2.2.2, we suggest that the level of transcription quality achieved for ST is sufficient and that no further revision by human transcribers is required for reaching the threshold proposed in 2.2.2. However, it is still true that almost one out of five symbols in the automatic transcription of SC is different from the consensus transcription. Therefore, more detailed analysis of the discrepancies is still in order. Additionally, if one would decide to employ human transcribers to verify and possibly improve all of the data, the analysis in the next stage can be used to indicate the phenomena that should receive special attention.

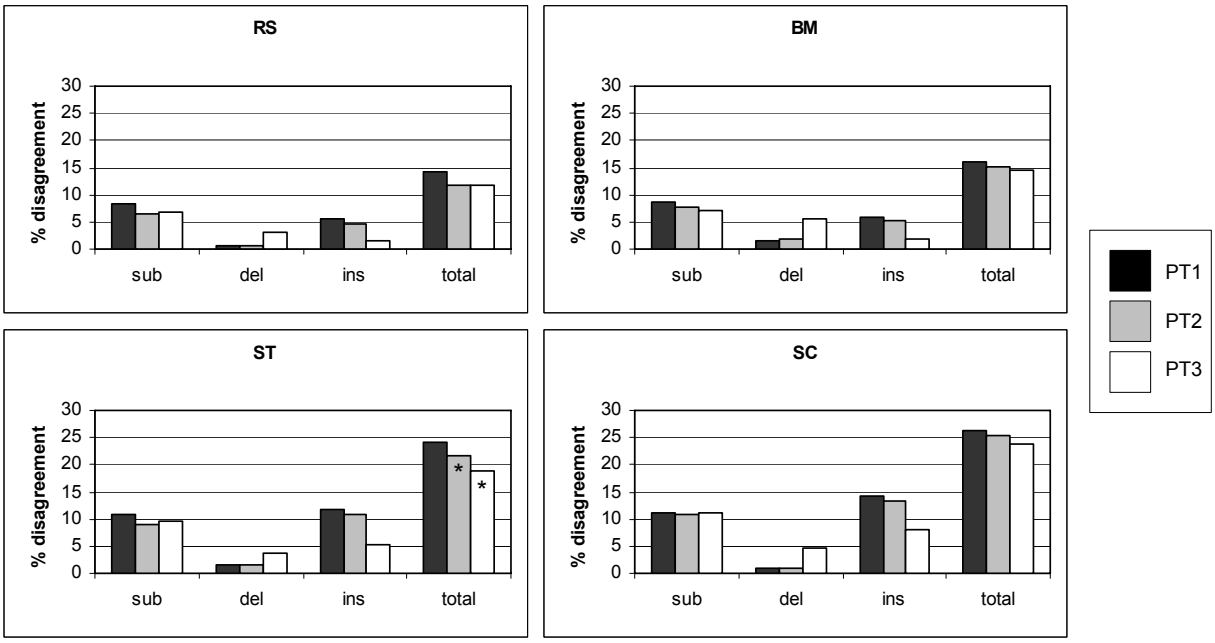

Figure 2-2 Summary of results of PT1, PT2 and PT3 for V-set. The asterisk (ST) indicates a significant difference in %disagreement of PT2 relative to PT1, and of PT3 relative to PT2.

Diagnosis – Stage 3.3

PT3 of the D-set was compared with the reference transcription. It shows the same trend as the results of the validation of PT3 of the V-set, except for RS, which shows a minor increase in percentage disagreement.

Table 2-11 Percentages disagreement of PT3 after diagnosis for all speech styles (D-set).

speech style	substitutions	deletions	insertions	total
RS	5.2	3.4	1.7	10.3
LC	7.2	4.1	4.1	15.4
IN	6.9	3.8	5.2	15.9
SC	9.2	4.8	6.8	20.8

A qualitative analysis was performed to determine which processes were responsible for the results. The most striking development is the increase of errors at word boundaries due to deletion of word-final /n/, /r/, and /t/. Many words in Dutch end with /@n/, so the deletion rule for /n/-deletion was applied very often. In addition, the ASR over-selected variants in which the segment was deleted.

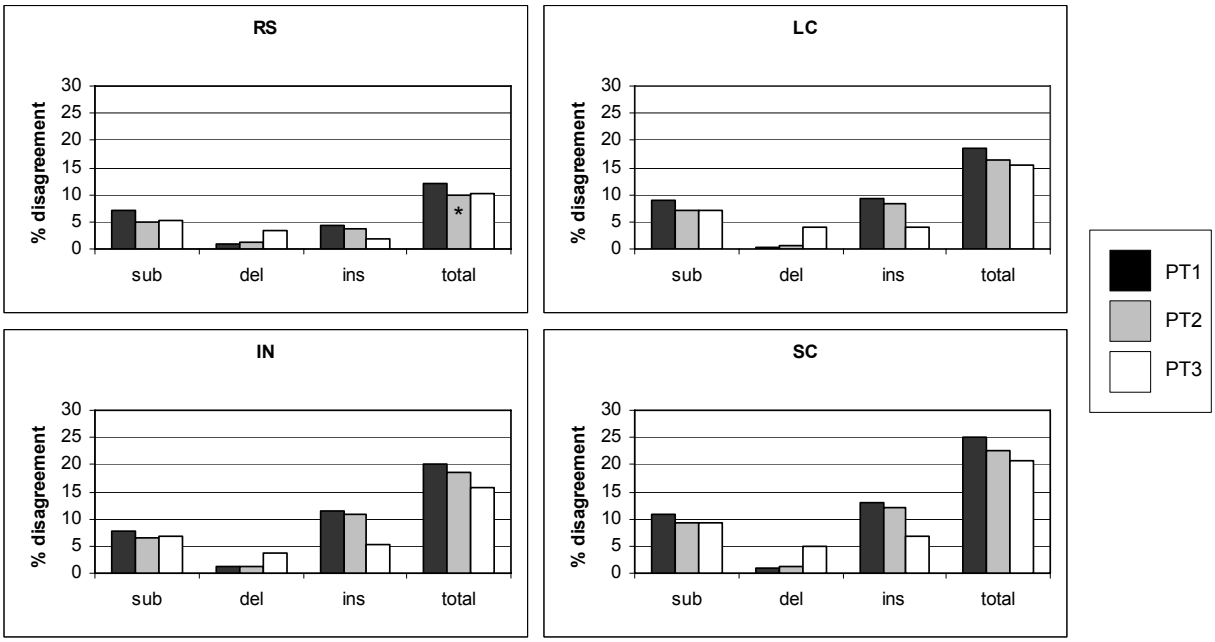

Figure 2-3 Summary of results of PT1, PT2 and PT3 for D-set. The asterisk (RS) indicates a significant difference in %disagreement of PT2 relative to PT1.

The distribution of substitutions varies between the four speech styles. For RS and SC, we found an increase in substitutions at word boundaries and a decrease in word-internal positions; for LC and IN, we found just the opposite. The rate of appropriate vowel reduction varies across the speech styles.

Remedies for the remaining errors will be defined in the next stage. The graphs in Figure 2-3 summarise the results obtained thus far for the transcriptions of the D-set. It can be seen that in the D-set there is only a significant improvement ($p \leq .01$) for RS from PT1 to PT2.

Remedy and decision – stage 3.4

A more detailed analysis of PT3 revealed that the majority of the deletions pertained to word-final /n/, /t/, and /r/, exactly those phonemes for which the ASR had to decide whether or not they had been realised. Too frequently, the ASR chose the pronunciation variants in which these phonemes were absent. Especially /t/ proved to be problematic. Apparently, the HMM recognizer that we used to determine the pronunciation variant that did fit the signal best was not able to reliably establish the presence or absence of a /t/. Most probably, this is due to the inherent difficulty to determine whether a short noise burst is (or is not) present. Still, the increase in deletion errors is compensated by a decrease in insertions of these phonemes (/n/, /t/, and /r/) in the total percentage disagreement.

With respect to the remaining substitutions, most still involve voicing discrepancies, but another prominent type concerns vowels. Distinguishing between voiced and unvoiced consonants in Dutch remains a difficult problem. For many speakers there is hardly a difference between voiced and unvoiced fricatives. This makes training models that reliably distinguish between /v/ and /f/, or between /z/ and /s/ extremely difficult. The situation with the difference between /G/ and /x/ is even more complex. Van de Velde (1996, p. 102-111) found some six different realizations of the velar fricative, and it seems that which variant is actually realized depends on the speaker, more than on anything else.

The rewrite rule for vowel reduction was applied to both multi-syllabic words and monosyllabic function words. It is mainly for the latter type that the ASR over-selected variants in which indeed vowel reduction was applied, both at word boundaries (‘ik’) and in word-internal (‘dat’) positions, while the full vowel was present in the reference transcription. It is well known that the acoustic properties of vowels overlap substantially. Therefore, there is no clear separation between full and reduced vowels. Apparently, the phoneticians who made the consensus transcription (and who did understand the words) tended to select the full vowel, unless the actual vowel was reduced substantially. Most probably, the tendency to prefer the full vowel is related to the canonical representation of

the word in the transcribers' mental lexicon. The HMM recognizer, on the other hand, could only rely on the acoustic properties of the vowels, since there was no built-in bias towards the canonical variant.

The analysis of the discrepancies between the consensus and the automatic transcriptions that remain after the first application of the third cycle suggests several ways in which the procedure for selecting the most appropriate pronunciation variant can be optimised. For one thing, it seems necessary to boost the prior probability of the canonical variants relative to reduced variants. This should help to reduce the spurious /t/-deletions and selection of reduced vowels. However, not all discrepancies can be tackled by means of straightforward tuning of the HMM recognizer. Specifically, the selection of voiced versus voiceless consonants will require other means.

For future research, it would be interesting to repeat the third cycle with improved versions of the ASR system. Perhaps, one might expect that further analysis of the remaining discrepancies between the consensus transcription and PT3 would have yielded additional rules for generating more style-specific pronunciation variants. However, recent research (Binnenpoorte et al., 2005) suggests that rule-based procedures will not be able to cover all pronunciation variants that occur in real-life speech, especially in the case of so-called multiword expressions.

2.5.4 Fourth cycle

Before entering the fourth cycle one should determine whether more improvements can be expected by 're-running' the third cycle. One should also decide whether an automatic transcription is already of sufficient quality for certain speaking styles, e.g. for ST and perhaps also for RS. Where human expertise is required, extra attention can be paid to the deficiencies in the automatic transcription identified in the diagnosis stage 3.3. As the purpose of the experiment was to demonstrate our procedure for the generation of automatic phonetic transcriptions, the fourth cycle, concerning human verification and correction, falls outside the scope of this paper. The role of human transcribers and the details of the transcription procedures in acquiring optimal transcription quality is analyzed elsewhere (Binnenpoorte et al., submitted).

2.6 Discussion and conclusions

In this paper, we have presented a bootstrap procedure for generating automatic phonetic transcriptions that has a twofold aim: a) obtaining sufficient quality automatic phonetic

transcriptions of LSCs and b) gaining new insights into phonological processes in real-life speech. We have also reported on a series of experiments that were intended to show how this procedure can be put into practice. The results of these experiments show that the proposed procedure is indeed effective for both goals.

First, the experiments show that the quality of the automatic transcriptions can indeed be improved without involving human transcribers for producing large amounts of transcriptions by applying the cycles of automatic transcription improvement proposed in this procedure. In addition, the results suggest that a high quality grapheme-to-phoneme converter together with some frequently applied phonological rules can generate transcriptions that are almost good enough for read speech and telephone conversations, given the criterion that the percentage disagreement with a consensus transcription may not be larger than the disagreement between that same consensus transcription and the transcription produced by an individual phonetician. However, 'good enough' for telephone conversations amounts to accepting that one out of five symbols in the 'cheap' transcription differs from the consensus transcription. This raises the question whether accurate transcriptions of large amounts of spontaneous speech are at all feasible.

The quality of the automatic transcription eventually achieved in our study ranges between 10.3% and 11.7% disagreement (D-set and V-set, respectively) with the reference transcription for read speech and between 20.8% and 23.9% disagreement (D-set and V-set, respectively) for spontaneous speech. At this point, it may be interesting to compare these figures to those obtained by other authors. For instance, Chang et al. (2000) report 20% disagreement for spontaneous American English and Saraçlar and Khudanpur (2000) report 26.6% disagreement on Switchboard data. Wesenick and Kipp (1996) report 11.6% disagreement for consonants in read aloud German. Thus, it seems that the procedure proposed in this paper yields results comparable to what is considered as state-of-the-art.

However, substantially lower disagreement levels were reported by Demuynck et al. (2004) in a study of automatic transcriptions for the Flemish part of the CGN corpus. These authors report 4.7%, 7.7%, 8.7%, 13.4% disagreement for RS, LC, IN and SC, respectively, measured between the automatic transcription and human-made transcriptions produced by individual transcribers. We believe that these superior agreement rates are due to a bias introduced by the procedures with which both the human and the automatic transcriptions were obtained: both were derived by 'correcting' a canonical transcription. As shown in Binnenpoorte et al. (submitted) such a procedure cannot but boost agreement rates. This assumption is supported by the explanation of the authors: the transcribers typically overlooked processes such as schwa insertion, homorganic glide insertion, and /n/-deletion due to nasal assimilation. These processes are well attested in continuous

spoken Dutch, but were not taken into account in the canonical example transcription presented to the transcribers. In addition, the transcribers failed to undo other processes that were actually applied in the canonical example transcription, such as syllable-final /n/-deletion. Therefore we can conclude that the disagreement figures express the degree of similarity between the automatic and the human-made transcription, but that the human reference transcriptions cannot be regarded as accurate representations of the actual speech signal.

An important result of the third cycle is that, despite all the improvements, discrepancies remain between the automatic transcription and the consensus transcription. A closer inspection of these discrepancies suggests possible improvements to this procedure. In general, for all four speech styles, the ASR tended to choose the shortest variant of a word in the lexicon. In spontaneous speech, this is less serious since a lot of reduction actually does take place. In any case, it seems that optimisation of the ASR, for example by tuning insertion penalties, or by retraining and reconfiguring the architecture of the acoustic models, such that shorter instances of phonemes are also recognised, may result in automatic transcriptions of better quality. Kessens and Strik (2004) have shown that reducing the number of states in the model topology of the /@/ led to higher agreement scores between automatic transcriptions and a reference transcription.

For a number of substitutions (e.g. voiced/voiceless, full vowel versus schwa) an ASR based on HMMs is probably not adequate. Here, a two-stage procedure is more appropriate. First, the ASR segments the speech. Then more detailed signal processing provides the basis for a final decision (cf. Truong et al., 2004). Similar approaches can also be applied to insertions and deletions: the HMM recogniser could generate a number of hypotheses for subsequent verification by additional processing.

Other discrepancies probably cannot be resolved with an approach like ours, which is crucially dependent on the ability to design rules that generate plausible pronunciation variants to be added to the lexicon. Some of the discrepancies observed in spontaneous speech are caused by extreme forms of reduction that lead to the deletion of entire syllables. For these phenomena, techniques based on the enumeration of pronunciation variants in a lexicon may be more appropriate. However, this would require a large manually verified, phonetically transcribed corpus for the extraction of the various pronunciation variants. Then, a sophisticated procedure to extract those plausible variants automatically from some corpus would have to be implemented (cf. Riley et al., 1999).

Our second objective was to obtain information about phonological processes in real-life speech. The results reveal that our procedure succeeded in providing new and more

complete information on the nature and frequency of various phonological processes in different speech styles. Booij (1995) describes the phonology of Dutch extensively, as does Ernestus (2000) for casual Dutch, but for automatic transcription more detailed information about the frequency of the various processes is indispensable. In our experiments, we used the consensus transcription of the D-set to derive this type of knowledge. For processes previously described in the literature, such as progressive and regressive voice assimilation, degemination, /n/-deletion, /r/-deletion, and /t/-deletion, we obtained specific information on their application frequency in the various speech styles. This information was then captured in rewrite rules subsequently used to produce an improved version of the automatic transcription. These rules were then validated against a consensus transcription of a second set of speech material, the V-set. Modelling the new-found knowledge clearly improved the automatic transcription. This provides extra evidence for the reliability and generalisability of the phonological knowledge extracted in our procedure. This type of validation is especially important when using a bootstrap procedure, which proceeds from observations of a relatively small set of speech data. Although the limited amount of data is dictated by the choice for a consensus transcription, the drawbacks of dealing with limited data sets should not be ignored. For instance, by extracting knowledge from a relatively small data set, one may risk tuning towards that specific small set of data. By using two independent sets, one for development and one for validation, we limited data-specific implementations.

Despite the time and effort spent in obtaining consensus transcriptions, we were still left with a limited amount of speech for discovering and testing phonetic knowledge that can be deployed in order to improve automatic phonetic transcriptions and that is also useful in linguistic research. This raises the question whether our approach is at all viable. The best way to answer this question is, probably, to analyze the results of the second cycle in more detail.

As was to be expected, it was found that a high quality grapheme-phoneme converter should account for common cross-word assimilation and degemination phenomena. The results also showed that the procedure proposed in this paper is able to re-discover virtually all phonological phenomena that are known to affect the phonetic surface forms. Yet, the analysis of the relative frequency with which these phenomena occur in different speech styles did provide new data. However, several phenomena were found that seem to be different between read speech and other, more extemporaneous, speech styles. One might object that the consensus sets were simply too small to allow for the detection of interesting novel phenomena. However, we are convinced that this is not the case. Common phenomena are sufficiently frequent to be detected in a small but representative sample,

while infrequent phenomena are mainly of theoretical interest. Therefore, the fact that our procedure is able to discover the important phenomena in a small set of consensus transcriptions in a well-analyzed and described language makes us confident that it will also able to do this for other languages for which detailed analyses and descriptions are not yet available. Therefore, the bootstrap procedure proposed in this paper is a useful addition to the tools and procedures that are now becoming available for generating automatic phonetic transcriptions of large speech corpora, especially corpora that contain substantial proportions of non-scripted speech.

In this paper we have exclusively relied on expert phoneticians to derive rules from the analysis of discrepancies between a consensus and an automatic transcription. Alternatively, one might want to apply machine learning techniques for the same purpose. Such data-driven techniques for the extraction of pronunciation rules were proposed by, among others, Cremelie and Martens (1997), Yang and Martens (2000a), Schiel (1999), and Kessens et al. (2003). The rules are extracted by means of a search for systematic mappings between a canonical transcription and some automatic phonetic transcription of a large number of utterances. Automatic phonetic transcriptions may, for example, be obtained as the output of a free phone recognizer, i.e., an ASR system that recognizes sequences of phones instead of sequences of words. Automatically derived rules can then be used to generate pronunciation variants. However, it should be evident that machine learning requires large amounts of data, much larger than the consensus transcriptions used in our approach. Moreover, it should also be clear that without the equivalent of consensus transcriptions it is not evident what the eventual quality of automatic transcriptions may be. Last but not least, the status of the automatically derived rules in terms of useful phonetic and phonological knowledge is somewhat unclear.

Rules obtained by means of a data-driven procedure can be represented in exactly the same manner as the rules derived by hand in the bootstrap procedure proposed in this paper. Moreover, it is quite likely that part of the data-derived rules will resemble the manually derived rules: in almost all cases these rules will be probabilistic, in the sense that they stipulate that a certain transformation applies in a certain proportion of the cases where its context is given. The most important advantage of data-derived rules is probably that the frequency estimates can be based on data sets that are much larger than can be afforded with consensus transcriptions. Because of the sheer size of the corpora that can be used for discovering relations between canonical or lexical forms and hypothetical phonetic transcriptions, data-driven techniques can in principle discover regularities that escape the analysis of discrepancies between automatic and consensus transcriptions. However, the pronunciation extraction process is less well supervised, making the status of less frequent

regularities unclear; the question arises whether they represent genuine phonetic processes, or whether they are the result of peculiarities of the ASR. As the number of less frequent regularities increases, it becomes increasingly difficult to interpret their status.

It has already been pointed out that there are reduction phenomena that are quite frequent; yet do not qualify as conventional rules, because the reductions are limited to the specific contexts of multiword expressions. It would seem that multiword expressions defy a general rule-based approach, irrespective of whether rules are derived by hand of by automatic procedures. For the time being, multiword expressions are most appropriately treated by adding them to the lexicon, with all plausible pronunciation variants observed in a sufficiently large and representative corpus.

There is at least one issue where the use of consensus transcriptions offers insights that are very difficult to obtain with data-driven approaches based on straightforward application of an HMM recognizer. These issues centre around the fundamental problems that were raised about the status of the voiced/unvoiced distinction for Dutch fricatives and plosives, and the distinction between full and reduced vowels. It is not evident how a fully automatic procedure could detect this kind of problems. Neither is it evident that an HMM machine is able to solve these transcription problems without taking recourse to other knowledge sources than straightforward statistics about local signal characteristics.

To conclude, our procedure for knowledge extraction is particularly suited when no phonetic transcriptions are available for an entire corpus of speech material and when little information is available on the languages in question. In the specific case of Dutch we obtained very useful information on the frequency of application of the various processes in the various speech styles, whereas with respect to the nature of the processes we mostly found confirmation of possible processes that had already been signalled. However, the increasing need for developing speech-based applications for languages that are less well described and documented, will require the rapid realization of basic resources such as speech corpora for these languages, often with limited financial means. In these situations a dual-purpose technique that appears to be effective both for obtaining automatic transcriptions of good quality and for extracting new systematic phonological knowledge from unexplored speech data may be extremely useful.

Acknowledgements

The authors would like to express their appreciation to T. Rietveld, W.H. Fletcher, and the expert transcribers for their contribution to this work. Furthermore, the authors would like to thank colleagues who read and commented on earlier versions of this paper.

VARIANT-BASED PRONUNCIATION VARIATION MODELLING FOR AUTOMATIC PHONETIC TRANSCRIPTION OF SPONTANEOUS SPEECH

CHAPTER 3

Reformatted from:

Diana Binnenpoorte, Catia Cucchiarini, Helmer Strik and Lou Boves (2004). Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

In this paper we present an experiment aimed at improving automatic phonetic transcription of Dutch spontaneous speech through a variant-based method of pronunciation variation modelling. For spontaneous speech, the literature does not always provide enough rules to describe its characteristic phonological processes. Therefore, other methods should be applied to model pronunciation variation for automatic phonetic transcription. We show that a large amount of manually transcribed phonetic data is an extremely useful source for collecting pronunciation variants and their prior probabilities. From the results we can conclude that the adopted method is indeed suitable for improving automatic transcription of spontaneous speech. Further improvements are expected to be obtained by combining this method with rule-based methods of pronunciation variation modelling.

3.1 Introduction

Annotated large speech databases are a rich resource for various linguistic studies. Manual annotation of speech signals is very time-consuming and costly. Especially phonetic transcriptions are known to be extremely labour intensive and therefore expensive. Recourse to automatic techniques would partly solve this problem. Although in the last decades considerable progress has been made in the field of speech recognition technologies, still an automatic speech recognizer (ASR) performs better on read speech than on conversational, spontaneous speech. This does not only apply to automatic speech recognition, but also to automatic transcription of speech (Cucchiarini & Strik, 2003).

However, many real-life situations in which ASR techniques can be applied concern spontaneous speech rather than read speech, which therefore constitutes a very good reason for trying to improve ASR performance on spontaneous speech. Since in this process automatic phonetic transcription has an important role to play, there are good reasons too for improving ASR performance on automatic phonetic transcription of speech data. This topic will be the focus of the present paper.

The fact that ASR performance on automatic transcription is systematically lower for spontaneous speech than for read speech can be explained in two different ways. The first explanation is that spontaneous speech is intrinsically more difficult to transcribe than read speech. The alternative explanation is that we are much better at modelling read speech than spontaneous speech, because the bulk of the knowledge accumulated so far in speech research does concern carefully pronounced laboratory speech, which is more similar to read speech than to spontaneous speech. The third possibility is a combination of the previous two: spontaneous speech is intrinsically more difficult to transcribe than read

speech, but the discrepancy in ASR performance on automatic transcription of read and spontaneous speech can be reduced by better modelling spontaneous speech.

Although we believe that spontaneous speech might somehow be more difficult to transcribe for both humans and machine, we are convinced that the current levels of ASR performance on automatic transcription of spontaneous speech can be improved to a certain extent through better modelling. In particular, current approaches to automatic transcription have made little use of the spontaneous speech corpora that are now becoming available for various languages, and which appear to be invaluable sources of information for various purposes, among which pronunciation variation modelling. In this paper we will show how automatic transcription of spontaneous speech can be improved by modelling some of the variation that characterizes this type of speech in a way that was not feasible until large spontaneous speech corpora became available: variant-based pronunciation variation modelling as opposed to rule-based pronunciation variation modelling.

In the remainder of this paper we go more deeply into the adopted method, and then we present the results after which a discussion is presented together with the conclusions.

3.2 Experiment

In the following section we first describe the method of the experiment, followed by a description of the speech material we used, how the automatic phonetic transcription is created based on a lexicon containing pronunciation variants, how a reference transcription of a small test corpus is made and finally how the latter was used to determine the quality of the automatically generated phonetic transcription.

3.2.1 Method

One way of obtaining automatic phonetic transcriptions is by having a speech recognizer in forced recognition mode select the variant that best matches the acoustic signal from a list of pronunciation variants contained in the lexicon. These variants can be generated in different ways (for an overview, see Strik & Cucchiarini, 1999). A very common method consists in generating the variants by means of rewrite rules that are either obtained from the literature or are extracted from speech data. A second option consists in extracting the variants directly from a large speech corpus (enumerated). The advantage of the first method, which we will call rule-based, is that the rules can be applied to all words in the lexicon, whereas in the second approach, which we will call variant-based, only variants that are found in the corpus can be included in the lexicon. However, the variant-based approach has the advantage that it allows modelling of word-specific phenomena that

cannot otherwise be captured by rules. Especially in spontaneous speech it often happens that highly frequent words undergo extreme reduction processes that can delete even up to complete syllables. Until recently, variant-based modelling could not be applied to Dutch, because we did not have an adequate corpus. Since we are now fortunate to have a large corpus of transcribed Dutch spontaneous speech, the Spoken Dutch Corpus (Oostdijk, 2000), we decided to study the effect of this type of pronunciation modelling on automatic transcription. In this experiment we limited ourselves to modelling frequently found words that are known to be enormously reduced in spontaneous speech.

3.2.2 Material

The speech material used in this experiment is divided into two parts, one for the extraction of variants to be added to the lexicon, the other for testing the performance of automatic transcription. The material was selected from the Spoken Dutch Corpus. We selected all the spontaneous material that had a manual phonetic transcription. This material consists of telephone conversations and dialogues (and multi-logues) that were recorded in home environments, using one central (stereo) microphone and a minidisk recorder. The different recording conditions of these two speech types result in different acoustic qualities. Nonetheless, we chose to use both types of spontaneous material because of the extemporaneous character of the speech that is almost the same in both conditions.

In Table 3-1 the most important statistics of the data are summarized: the total duration of the speech material, the number of words, the number of unique words and the average number of pronunciation variants per word.

Table 3-1 Statistics of both train and test set.

	duration (hh:mm:ss)	#words	#unique words	average #variants
TRAIN	24:26:07	304,502	14,113	2.2
TEST	0:13:04	2,822	676	1.7

In total 7620 words in the training set were found with only one pronunciation, most of which are proper names, infrequent inflections of verbs and broken words (start-repairs). The forty most frequent words cover 50% of all the words in the training set and most of these are short (monosyllabic) function words and first person inflections of the verbs ‘zijn’ (‘*to be*’) and ‘hebben’ (‘*to have*’). Multisyllabic function words, such as ‘natuurlijk’ (‘*of course*’), ‘helemaal’ (‘*totally*’), ‘eigenlijk’ (‘*actually*’) and ‘allemaal’ (‘*all*’), are also very frequent and can be found in the top hundred of most frequent words.

3.2.3 Lexicon training set

The broad phonetic transcriptions were obtained by having trained transcribers verify and possibly correct an optimized automatically generated phonetic transcription. Then, in a second round, the resulting transcriptions were verified and corrected, if needed, by another transcriber. Besides this manual phonetic transcription and the original orthographic transcription, the training material is also manually time-aligned to the speech signal on word level. Thus, every orthographic entity is unambiguously linked to a phonetic transcription.

All the word types in the training set are collected together with their transcription and are sorted on frequency. Then a prior probability for each pronunciation variant is calculated given the frequency of occurrence of its orthographic counterpart in the training material. The list created this way contains all possible pronunciations of the words found in the training set and their probability of occurrence. 80 orthographic words from the test set did not occur in the training set. For these words a unique canonical phonetic transcription was obtained, by consulting the general CGN lexicon and these transcriptions were assigned a prior probability of 1. Furthermore, 65 words in the test set only occurred once in the training set and were assigned the observed pronunciation variant in the lexicon.

3.2.4 Automatically generated transcription - AGT

We used an ASR (Strik et al., 1996) in forced recognition mode to choose the most likely pronunciation variant from the lexicon given a class-based language model and the acoustics of the speech signal. The acoustic models are continuous density hidden Markov models with 32 Gaussians per state trained on phonetically rich sentences uttered through a telephone. We converted the wide band material of the test set, the recordings in the home environments, to telephone bandwidth in order to avoid the mismatch between the acoustic properties of the models and the test data.

For each utterance in the test set a pronunciation lexicon is extracted from the training lexicon, where each word in the utterance has all the pronunciation variants as they were found in the training material. The language model was a class-based bigram model. The prior probabilities of the pronunciation variants of a word are captured in the unigram part. Here, the classes, or categories, are the words in the utterance; the transitions between words are modelled by the class bigram (Brown et al., 1992). The result of the forced recognition is a sequence of the pronunciation variants of the words in the utterance that best matches the speech signal, the AGT.

3.2.5 Reference transcription – RT

A reference transcription (RT) can serve as a benchmark against which other transcriptions, in this case an AGT, can be validated. A consensus transcription is probably the best possible approximation of the 'true' transcription (Shriberg, 1991). Two phonetically trained and experienced listeners were asked to make a consensus transcription of the speech material in the test set. They transcribed from scratch and had to agree on each symbol in the transcription. They used the same symbol set as was used for the AGT. This led to a broad phonetic consensus transcription, which will serve as the RT in this experiment.

3.2.6 Alignment

A dynamic programming algorithm was used to make an alignment between the AGT and the RT in order to determine the agreement between the former and the latter. The program provides the number of substitutions, deletions and insertions on phoneme level. Each of these errors is assigned a weighting, which is used as a distance measure during the alignment procedure. The weightings are calculated in terms of articulatory features, such as place and manner of articulation, voice, lip rounding, length, etc. The results of the alignment show in what respects the AGT differs from the RT.

3.3 Results

3.3.1 Phone error rates

In the first row in Table 3-2 the results of the alignment between the AGT and the RT are shown in percentages of substitutions, deletions and insertions on phoneme level. The total percentage disagreement (last column) is the phone error rate (PER). In order to put the data in perspective, the second row gives the result that was obtained by modelling frequent phonological processes by means of rules for the same data (Binnenpoorte & Cucchiarini, 2003). Finally, in the last row the percentage disagreement on phoneme level between a simple concatenation of canonical forms and the RT for the same material is displayed.

Table 3-2 Quantitative results of alignment between AGT and RT and previously found results.

%	substitutions	deletions	insertions	total
AGT	10.01	7.22	4.50	21.73
STATIC	10.37	1.57	11.83	23.77
CANON	12.50	2.00	12.87	27.37

In Binnenpoorte et al. (2003) four trained transcribers were asked to transcribe a part of the spontaneous speech material as contained in the test set. When comparing their transcriptions with the corresponding part in the RT we found total PERs ranging from 13.4% to 15.7%, with inter-transcriber agreement ranging from 85.7% to 94.9% (where the latter figure relates to agreement found by comparing the transcription of the first transcriber with the correction of that first transcription by a second transcriber). Although the data set of the human transcription differs from the AGT, the results obtained in this experiment surpass the best AGT performance in previous experiments. Still the AGT does not come close to human performance yet, which is not surprising if we consider that in this experiment we only applied the variant-based method.

3.3.2 Analysis of PERs

Closer inspection of the output of the alignment between the AGT and RT reveals that for all substitutions, deletions and insertions a relatively small number of phonetic processes cover more than half of the errors. To illustrate, the 13 most frequent substitutions (8.3% of all the substitution types) are responsible for 50% of the substitution errors. In case of the deletions, 50% of the errors can be accounted for by only 4 deletion types (11.4% of total), and also the 4 most frequent insertions (12.1% of total) are responsible for 50% of the insertion errors.

The most frequent substitutions are confusions between phonemes that only differ in one articulatory feature, see Table 3-3, primarily related to the feature voice (in fricatives and plosives) and length (in vowels). In addition, confusions between any vowel and schwa are also frequent. Most deletions are related to /@/, /r/, /d/ and /n/. Finally, for insertions we found that most of the errors are due to insertion of /@/, /n/, /r/ and /t/.

Table 3-3 Top five of substitutions, deletions and insertions in dataset containing 8063 phonemes.

substitutions		deletions		insertions	
#	phones	#	phones	#	phones
51	/G-x/	117	/@/	68	/@/
50	/s-z/	68	/r/	60	/n/
46	/d-t/	63	/d/	39	/r/
41	/A-@/	47	/n/	29	/t/
37	/f-v/	39	/t/	24	/j/

3.4 Discussion

The data in Table 3-2 show that our attempt to optimise automatic phonetic transcription by means of a lexicon with pronunciation variants observed in a large manually transcribed corpus has been successful. The improvements are mainly the result of fewer insertions, which means that the ASR has chosen variants in which reduction of specific phonemes was modelled. On the other hand, the number of deletions has risen enormously. We believe that many –but not all– of the remaining discrepancies between our APT and RT are due to inherent limitations of the HMM recogniser used as a transcription tool. The 117 /@/ deletions can illustrate this: The topology of the acoustic models in our ASR requires that phonemes span at least 30 ms to be detected. It seems that the two expert listeners had a lower durational threshold for /@/. We believe that we see similar problems with the other frequent insertions and deletions. Dutch has a substantial number of frequent unstressed syllables with a vowel followed by /r/ or /n/. In all these cases the acoustic basis for the detection of the individual phonemes in the canonical representation is rather weak, especially in spontaneous speech. More often than not, the presence of one 'sound' is fully encoded in the phonetic details of its neighbours. Phoneticians are able to reach a high degree of agreement on the segmental transcription of these syllables (cf. the agreement data in Goddijn & Binnenpoorte, 2003), but this is probably due to a common interpretation of these acoustic complexes, biased by the fact that they understand the words and therefore can rely on knowledge of the underlying canonical form. However, a phone-based HMM system is fundamentally unable to reproduce this behaviour.

The most frequent substitutions that remain in our approach are related to the feature 'voice'. Due to the fact that the lexicon only contained observed pronunciation variants, we may have missed a number of realistic variants, especially in words that are not among the

most frequent. Also, our approach may not be the best solution for cross-word voice assimilation, a process that is known to be quite important (Binnenpoorte & Cucchiarini, 2003). However, also in this case we think that the HMM system is partially to be blamed. Especially for fricatives 'voice' has quite an uncertain status. As a consequence, it is virtually impossible to train HMMs that can tell the voiced and unvoiced cognates apart. To approximate human-like performance in voiced-unvoiced distinction we will need a two stage procedures that operates on the segmentation of the HMM system, and that applies independent acoustic evidence for the classification.

In this paper, we adopted a variant-based approach to generate pronunciation variants. We put all observed variants in the lexicon. A disadvantage of this approach is that only 'seen' variants of a word can be modelled. For words that did not occur in the corpus from which the variants were derived, the lexicon will contain only the canonical form. In our case, 1.4% of the total number of discrepancies between APT and RT originates from the 80 'unseen' words. To obtain pronunciation variants for these and other less frequent words we can use the manually annotated corpus for the extraction of rules. This can be achieved by comparing the manually transcribed data with canonical transcriptions of that same data to generalize over all differences given a certain context (Riley et al., 1999; Scharenborg & Boves, 2002; Wester, 2003).

The combination of rewrite rules together with prior probabilities of pronunciation variants could be especially promising for multiword expressions. These are frequently used expressions in everyday language, such as institutionalized phrases. Most of the time, the individual words of a multiword expression are pronounced with much more reduction in the multiword construction than in other, less frequent, constructions. Multiword expression should therefore be considered as one entity in the same way as 'normal' words.

3.5 General discussion

In this paper we have shown that automatic phonetic transcription of spontaneous speech can be improved to a certain extent by modelling pronunciation variation through a variant-based method which could not be applied before a large corpus of spontaneous speech became available for Dutch. It is clear that the more transcribed data are available, the better spontaneous speech can be modelled, which, in turn, means that the APT can be improved such that more transcriptions can become available at lower costs.

In spite of this enhancement in performance, there is still much room for improvement to obtain performance levels that much more resemble those obtained for read speech. However, this is not surprising if we consider that in this experiment only the variant-based method of pronunciation variation modelling was applied, thus neglecting the modelling of

other processes that, as we know, are best addressed through the rule-based method. The challenge will now be to find the optimal combination of these two methods which provides the best performance levels. This will be the focus of our research in the near future.

Based on the results of the experiment reported on in this paper we can conclude that the adopted technique of modelling real-life pronunciation variants does improve automatic phonetic transcription quality, but is still not sufficient to resemble human phonetic transcriptions. A combination of variant-based and rule-based methods will probably offer the best solution.

Acknowledgements

We would like to thank J. Sturm and O. Scharenborg for their contributions to this research.

MEASURING PHONETIC TRANSCRIPTION QUALITY IN LARGE SPEECH CORPORA

CHAPTER 4

Reformatted from:

Diana Binnenpoorte, Catia Cucchiarini and Lou Boves. Measuring phonetic transcription quality in large speech corpora. Submitted to *Language Resources and Evaluation*.

In this paper we investigate the widely used method for obtaining manual phonetic transcriptions in large speech corpora in which a predefined example transcription is verified and corrected by human transcribers. This procedure saves time and money compared to transcription from scratch. In evaluating this procedure, transcription quality is usually established by measuring inter-transcriber agreement. We argue that this is not a suitable measure, since first, similarity between symbols does not imply correct use of symbols, and second, percentages of agreement are artificially inflated because of the bias effect of the example transcription. Therefore, we introduce an additional measure to establish transcription quality. In addition, we propose a more detailed analysis of the resulting transcriptions that goes beyond the percentages of agreement to reveal the actual underlying processes.

4.1 Introduction

Large speech corpora are extremely useful for linguistic research and application development. In the last few years many large corpora have been compiled with several different purposes in mind, such as Buckeye (Pitt et al, 2005), Corpus of Spontaneous Japanese (CSJ) (Furui et al, 2000), Spoken Afrikaans Language Resource (SALAR) (Wissing et al, 2004), Switchboard (Godfrey et al, 1992), and Verbmobil (Hess et al, 1995). These corpora contain considerable amounts of non-scripted speech, such as telephone conversations, spontaneous human-human interactions, and human-machine interactions, in order to satisfy the general need for real-life speech data in advanced research in linguistics, phonetics and speech technology. However, before these corpora can become useful for research and application purposes, the speech needs to be annotated at various levels, depending on the goals of the research or application. The above mentioned corpora all come with a phonetic transcription of parts of the material, either as originally planned or as additional annotations provided after the completion of the corpus. Since the sizes of these corpora are such that a complete manual phonetic transcription by experts is practically impossible, given the usual restriction of time and money, procedures have been developed to generate phonetic transcriptions as efficiently as possible. In all the above mentioned corpora, human transcribers had to correct a given example transcription; in some cases that was just a concatenation of canonical forms of the words in the orthographic transcription; in other cases an optimised automatic transcription was provided as point of departure. A similar transcription procedure was adopted in the Spoken Dutch Corpus, CGN (Oostdijk, 2002). The CGN was finished in 2004 and contains about 9 million words of contemporary Dutch as spoken in the Netherlands and Flanders. It

is intended to be a multi-purpose corpus that serves the needs of a broad range of researchers and application developers in the field of linguistics and speech technology. About 1 million words (the *core corpus*) were enriched with manually verified broad phonetic transcriptions. For the production of the phonetic transcription in the *core corpus* of the CGN, transcribers were asked to verify and correct an example transcription in a broad phonetic symbol set: Dutch SAMPA tailored to CGN (Gillis, 2001). Although transcription time and costs can be reduced by following such a procedure, it is unclear if and how the example transcriptions affect the quality of the transcriptions that are ultimately attained.

The quality of manual phonetic transcriptions of large amounts of speech data is often evaluated by measuring inter-transcriber agreement (Eisen, 1993; Greenberg, 1996; Kikuchi & Maekawa, 2003; Pitt et al., 2005; Raymond et al., 2002; Wesenick & Kipp, 1996). For that purpose, two or more transcriptions of the same material made by different transcribers are compared. The agreement between human transcriptions serves as a measure for the quality obtained with a specific procedure, since it is assumed that high agreement scores indicate high transcription quality. However, the use of inter-transcriber agreement as a measure of transcription quality raises questions.

First, it is questionable whether one can rely on agreement scores as suitable indications of transcription quality. For instance, Pye et al., 1988, p. 19, observe that "For most of the available studies that report a percentage of transcriber agreement, the number given is typically greater than 85%" and "Although such figures may allay concerns about the integrity of the resultant transcript, there is little objective foundation for placing confidence in the number". Among the reasons that these authors mention for questioning percentages of transcriber agreement are the lack of information about the level of transcription detail and the details of the comparison procedure and the limited number of comparisons. After all, the degree of agreement is heavily dependent on the level of transcription and on the number of comparisons, where agreement between two transcriptions tends to be higher than agreement measured on multiple comparisons of transcriptions. The question about the value of inter-transcriber agreement as a measure of quality becomes even more urgent if transcriptions are the result of checking and correcting example transcriptions. It is to be expected that transcribers can focus on a limited number of phenomena when making transcriptions. As a consequence, when they have to edit an example transcription, they will probably leave the example transcription intact for the phenomena that they do not concentrate on. This scenario seems even more plausible if we consider the time pressure under which transcribers usually have to work. Experience in the CGN project has confirmed that transcribers may fail to notice several, even frequently

occurring, phonological processes in correcting an example transcription (Demuynck et al., 2004). The more transcribers leave the example transcription intact, the higher inter-transcriber agreements will be. A biasing effect from the example transcription might result in high inter-transcriber agreement indices. High agreement scores that are due to the fact that the transcribers leave the example transcription intact, do not necessarily indicate that the transcriptions are accurate representations of the speech signal.

Second, percentage inter-transcriber agreement only shows a part of the full picture. Percentage agreement scores are summary figures in which the underlying processes that affect the ultimate score cannot be completely expressed. Two different transcription pairs (e.g. A-B and C-D) can have equal agreement scores that are based on different phenomena or symbols. For example, the total number of differences between a transcription pair (A-B) can be composed of 10 deletions, whereas the same number of differences between the other transcription pair (C-D) can be composed of 10 substitutions. In both cases the same percentage agreement score will be reported, regardless of the differences between the data from which the scores are obtained. Furthermore, if percentage agreement is computed on the basis of symbols, no distinction is made between subtle and more conspicuous differences; only the proportion of symbols that deviate is ultimately reported. In the case of substitutions, for example, both substitutions between similar phonemes and between more different ones are treated as equally serious. And finally, transcriber-specific phenomena cannot be revealed either without a more detailed analysis of the underlying differences.

In this paper we aim at evaluating the transcription procedure in which transcribers have to verify and correct a given example transcription. We will do this by analysing transcriptions produced by the transcribers who were employed for the phonetic transcription of the Northern Dutch part in the CGN project. For measuring transcription quality two measures will be adopted: first, the commonly used inter-transcriber agreement and, second, an alternative measure, viz., agreement between individual transcribers and a reference transcription that, at least in part, remedies the problems with inter-transcriber agreement as a measure of transcription quality. In the second part of this study, we perform a detailed analysis of the contents of the transcriptions. It will be shown that this yields valuable information that cannot be conveyed by reporting only agreement figures.

In the remainder of this paper we will discuss the issues concerning the measurement of transcription quality in section 4.2. In sections 4.3 and 4.4 transcription quality measurements are obtained by a) measuring inter-transcriber agreement and b) applying our additional measure. In section 4.5 the qualitative analysis is presented after which the

results and implications for future transcription projects are discussed in section 4.6. Finally, in section 4.7 we will present some final conclusions.

4.2 Measuring transcription quality

Assessing phonetic transcriptions is not straightforward. The notion of transcription quality has two different aspects, validity and reliability. Phonetic transcriptions can be viewed as representations or 'measurements' of the speech signal. The validity of a transcription expresses to what extent the measurement indeed measures what it is supposed to measure, i.e. to what extent the symbolic notation actually reflects the speech signal (Cucchiarini, 1993). Transcription validity can be established by comparing a phonetic transcription with the "true" reference transcription, which is a precise reflection of the speech signal. However, a unique reference transcription of a speech signal does not exist, because a string of symbols contains less information than the original speech signal. Therefore, validity of phonetic transcriptions can only be approximated – perhaps closely.

Reliability of phonetic transcriptions can be expressed in terms of the degree of consistency observed between repeated 'measurements' of the same speech signal (Ball & Rahilly, 2002; Cucchiarini, 1993; Cucchiarini, 1996; Tinsley & Weiss, 1975). Reliability, or consistency, can be determined on both intra- and inter-transcriber level. Consistency in a transcription is a prerequisite for validity; inconsistent use of phonetic symbols for one and the same speech signal can never yield a valid representation of that specific speech signal. If multiple independent consistency measures can be obtained for the same data, consistency can be used as an estimate of validity. After all, if a speech signal is labelled with the same symbol 100 times by independent transcribers, that symbol is likely to be the correct one, i.e. a valid representation of the speech signal.

For consistency to be a good approximation of validity, many independent repetitive observations must be made. This implies that many transcriptions should be collected of the same speech material such that these can be compared. In Eisen (1993), Greenberg et al. (1996), Kikuchi and Maekawa (2003), Raymond et al. (2002), and Wesenick and Kipp (1996) consistency is established by measuring inter-transcriber agreement for two to eight different transcribers. In actual practice, inter-transcriber agreement must be estimated on a small sample of speech that is processed by all transcribers who worked in a corpus production project. Therefore, the measure is dependent on the amount of speech processed by all transcribers, and also on the number of transcribers.

In the framework of transcription evaluation in large speech corpora, consistency as a substitute for validity has an additional limitation; viz., the fact that in most of the large

speech corpora the transcriptions are produced by editing an example transcription. Maximum consistency will be reached if the example transcription is left intact. Thus, high consistency does not necessarily mean that the symbols in the transcriptions are indeed valid representations of the speech signal. Without examining the number of symbols that were actually changed in the example transcription, percentage agreement is difficult to interpret. In general, inter-transcriber agreement scores obtained by editing an example transcription are likely to yield too optimistic an estimate of transcription quality.

Although a true criterion transcription does not exist to measure validity, it is possible to approximate the ground truth. Individual expert transcriptions are known to be subjective and contain idiosyncratic elements (Cucchiarini, 1993) and are therefore unsuitable as true reference transcription. We believe that a consensus transcription - as suggested by Shriberg et al, 1984 – obtained through a procedure in which a group of transcribers discusses to reach agreement on each symbol contained in the transcript comes closest to the ideal reference transcription. The mutual agreement of the group of transcribers on each symbol is the result of negotiation such that both subjectivity and idiosyncrasy are minimized. The consensus transcription we suggest to be the reference transcription should be generated from scratch, to prevent as much as possible any form of bias of any given example transcription. The comparison between an individual transcription and such a consensus transcription better indicates to what extent the symbols in the individual transcription reflect the speech signal.

Having defined the additional measure, we can proceed to examine the suitability of inter-transcriber agreement for transcription evaluation in large speech corpora. A comparison of the results obtained after applying both measures can reveal whether the given example transcription has a biasing effect on inter-transcriber agreement scores.

4.3 Experimental setup

In this section we describe the procedure that we used to analyse the transcriptions produced by the transcribers employed in the Northern Dutch part of the CGN project.

4.3.1 Speech material

The speech material on which we conducted our experiment was taken from the Northern Dutch variety of CGN. The sample was composed of 16 different one-minute fragments, representing four broad categories of speech styles: read speech (RS), lectures (LC),

interviews (IN) and spontaneous conversations (SC). Twenty different speakers were involved in the recordings, eleven male and nine female, who came from various regions in the Netherlands. In Table 4-1 some quantitative data on the sample is given.

Table 4-1 Statistics of the speech sample taken from the CGN.

	duration (mm:ss)	#utterances	#words	#phonemes
RS	04:57	141	689	2789
LC	05:09	120	912	3243
IN	03:01	83	523	1718
SC	03:01	100	615	1777
total	16:08	444	2739	9527

4.3.2 Transcriptions

Three types of transcriptions were generated for the experiment. First, we needed a reference transcription, which we defined as a consensus transcription. Second, we generated the example transcription that was to be edited by the individual transcribers. And third, four individual transcriptions were made of the material following the CGN procedure.

Consensus transcription

Two highly trained transcribers from the Netherlands made the consensus transcription of the speech material in the sample. During the transcription process they sat together in a quiet room. They transcribed from scratch (no example transcription was at their disposal) and had to agree on each symbol included in the transcript. The symbol set they used is the same as the one used for the CGN transcriptions (Oostdijk, 2004a). The original orthographic transcription was available and could be consulted in case of doubt. Making a consensus transcription is a time-consuming and costly enterprise; on average, it took about 60 minutes to transcribe one minute of speech. Once the consensus transcription was finished, it was aligned with the original orthographic transcription. If the phonetic transcription implied the deletion or insertion of a word (in all cases short function words that can be reduced to less than a single speech sound) the orthographic transcription was adapted. The corrected orthographic transcription then served as the starting point for the remaining transcriptions in the experiment.

Example transcription

The example transcription that served as a starting point for the individual transcribers was obtained by concatenating the citation forms of the words as found in the CGN lexicon (Oostdijk, 2004b). In these canonical transcriptions all obligatory word internal phonological processes are applied, whereas optional word internal processes are not; therefore, the phonetic transcriptions in the CGN lexicon represent very carefully pronounced citation forms. Some optional highly frequent word boundary processes, such as assimilation of voice and degemination, were applied to the concatenated transcription by means of conditional rewrite rules; more details can be found in Binnenpoorte & Cucchiarini (2003).

Individual transcriptions

The manual transcriptions made for the experiments were produced in exactly the same way as in the CGN project. In total eight transcribers were employed for the phonetic transcription during the project. At the moment of the experiment five transcribers were active, who were asked to participate in the experiment. Four of them succeeded in completing the task. These four transcribers (indicated as IT1, IT2, IT3, and IT4 in the remainder of this paper) had been working on the Northern Dutch CGN transcriptions for more than five months at the moment of the experiment. They can be considered as representative, since these four transcribers generated almost three-quarters of the phonetic transcriptions of the Northern Dutch part of the corpus. As in the CGN project, their task for the experiment was to make an auditory transcription by correcting an automatic transcription of the sample material according to a strict protocol (Gillis, 2001). The protocol states that the given transcription must be adapted by substituting, inserting and deleting phonemes when a different sound is perceived than the one reflected in the example transcription. Transcribers were explicitly instructed to only change the example transcription in case they were confident that it was incorrect. Transcriptions in the CGN corpus are phonemic. Therefore, gradual processes like degree of voicing in plosives and fricatives and monophthongization or diphthongization in vowels cannot and need not be expressed. According to the protocol loan vowels and nasalized vowels are only allowed in the transcription of loan words. Another important restriction lies in the requirement that a one-to-one relation between the orthographic and the phonemic transcription had to be maintained at the level of the words. Since the example transcription is based on the orthography, no words were allowed to be deleted or inserted. Therefore, each word in the orthographic transcription must be represented by at least one symbol in the phonetic transcription, and the phonetic transcription cannot contain symbols that cannot be

accounted by a word in the orthographic transcription. As in the CGN, the fragments containing spontaneous conversations (SC) were corrected twice. A second transcriber corrected the transcription of the first one. The rationale behind this double-check of the transcriptions of spontaneous conversations is the hypothesis that this type of speech is intrinsically more difficult to transcribe. The double-check procedure will probably result in more accurate transcriptions. In the context of the experiments described below it means that we do not have four independent transcriptions of the spontaneous speech items.

4.3.3 Comparing transcriptions

The individual transcriptions are evaluated in two different ways, by: a) measuring inter-transcriber agreement, and b) measuring the deviation from the reference transcription. In Table 4-2 the comparison scheme is presented. Each individual transcription (IT) is compared with each of the three other transcriptions and the reference transcription (REF).

Table 4-2 Comparison scheme to obtain inter-individual transcription (dis)agreement.

	IT1	IT2	IT3	IT4	REF
IT1		√	√	√	√
IT2			√	√	√
IT3				√	√
IT4					√

All comparisons were performed individually for each speech style considered in our experiment; RS, LC, IN, and SC, see section 4.3.1.

A dynamic programming algorithm, Align, as proposed in Cucchiarini (1996), is used to align two phoneme strings and to determine the distance between these strings by comparing them on a phone-by-phone basis. During the alignment process, costs for deletions, insertions and substitutions are calculated on the basis of the articulatory features of the sounds represented by the symbols. Align does not allow substitutions between consonants and vowels, because the two major phoneme categories are specified with different feature sets. In addition to an alignment score, Align also outputs the number of substitutions (S), deletions (D) and insertions (I), which is subsequently used to calculate the percentage disagreement:

$$\%disagreement = \frac{\#S+\#D+\#I}{\#symbols} * 100\%$$

In this formula the number of substitutions, deletions and insertions is related to the number of phonemes present in a reference transcription, or norm, to which the other

phoneme string is compared. In case of inter-transcriber agreement none of the two ITs is more suitable to be a norm transcription than the other. Therefore, we calculated percentage disagreement as the average of two disagreement scores, in which the number of phonemes in both ITs each served as a reference to which the number of substitutions, deletions, and insertions is related.

In the literature, inter-transcriber comparison measurements are usually expressed in percentage agreement. We prefer to present the distances between individual transcribers in percentages disagreement (which is simply 100% minus percentage agreement) because these numbers are easier to compare. Besides the numerical output, *Align* also yields information about the nature of the discrepancies between two transcriptions which were used for a more detailed analysis.

4.4 Results

4.4.1 Inter-transcriber disagreement

In total six transcription pairs were compared for each speech style using the *Align* program. In Table 4-3 the results are displayed for all transcription pairs in all four speech styles.

Table 4-3 Percentages disagreement measured between pairs of individual transcriptions for all speech styles.

	RS	LC	IN	SC
IT1 – IT2	6.2	10.7	12.2	14.0
IT1 – IT3	4.8	9.4	11.1	9.6
IT1 – IT4	5.1	10.2	11.9	14.0
IT2 – IT3	4.4	8.0	8.9	11.4
IT2 – IT4	3.8	8.2	8.5	4.9
IT3 – IT4	3.7	7.7	8.1	11.6

From Table 4-3 it can be observed that the percentage disagreement between ITs is considerably lower for read speech (RS) than for the three other speech styles. It seems as if RS is easier to transcribe, which can be explained by the fact that the pronunciations of words in RS are closer to their canonical representations. Human transcribers possibly make use of this fact by either their internal mental (canonically based) lexicon or, as in the CGN case, by the given example transcription that was based on canonical representations in the CGN lexicon. In our experiment, the given transcription is already close to what was

actually pronounced, such that the transcribers needed to change fewer symbols in the RS sample as opposed to the other styles, so that they could concentrate on fewer pronunciation phenomena.

The double-check principle that was adopted for the SC fragments clearly has an effect on inter-transcriber disagreement scores. In our experiment IT3 corrected IT1's transcriptions and IT2 corrected IT4's transcriptions; consequently the percentages disagreement between IT1 and IT3 on the one hand, and between IT2 and IT4 on the other are considerably lower compared to the other inter-transcriber disagreement percentages obtained for SC. Correcting other transcribers' transcription can be considered as editing a highly optimised example transcription.

The transcriber pair IT1-IT2 shows most disagreement in all speech styles compared to other pairs, while IT3 and IT4 seem to agree most. Transcriber pairs in which IT1 is involved show higher disagreement percentages than other transcriber pairs. Whether there is an underlying cause for these observations will be revealed in the following analyses.

4.4.2 Symbols changed in example transcription

Since the transcriptions were produced by editing an example transcription, it is possible that this example transcription had an influence on the transcriptions that were obtained. As a consequence, the agreement between the transcription may be biased by the given example transcription. Owing to time pressure and the difficulty to concentrate on many phenomena at the same time, transcribers may tend to leave the example transcription intact at some points, despite the fact that it does not represent the speech signal accurately. Therefore, we calculated the number of symbols that were changed in the example transcription; see Table 4-4.

Table 4-4 Percentages of symbols in the example transcription that were modified by ITs.

	RS	LC	IN	SC
IT1	11.3	16.5	19.9	23.4
IT2	9.4	12.1	14.3	22.1
IT3	11.2	15.5	17.3	22.9
IT4	10.2	14.0	16.8	21.4

For comparability reasons, the percentages of symbols changed in the example transcriptions for SC were also calculated for IT2 and IT3 although they started from IT4 and IT1, respectively. We simply calculated the number of symbols that were different

between the example transcription (that was never seen by IT2 and IT3) and the transcriptions of IT2 and IT3. The percentage of symbols that were actually changed by IT2 and IT3 during the double-check procedure, are precisely the scores for disagreement in Table 4-3 between IT2-IT4 and IT1-IT3 (4.9% and 9.5% respectively). IT2 and IT3 corrected IT4 and IT1 respectively for the SC fragments. If during the double-check phase all changes made by IT2 (and IT3) were indeed further corrections, then the percentage of symbols differing between the example transcription and the resulting transcriptions of IT2 would be close to a sum of the percentage of symbols IT4 first changed in the example transcription and the percentages of symbols IT2 changed in IT4's transcription (the percentage disagreement between IT2 and IT4, 4.9%). Table 4-4 shows that this is not the case, which implies that IT2 changed a number of symbols back to the initial example transcription, which IT2 never saw. The same holds for IT3 in relation to IT1. Furthermore, the figures in Table 4-4 reveal that the ITs made approximately the same number of changes to the example transcription, and that the proportion of symbols changed increases as the speech style becomes more spontaneous. However, from a comparison of the inter-transcriber disagreement percentages in Table 4-3 and the total proportion of changes in Table 4-4 it can be seen that the ITs changed different symbols in the example transcription.

4.4.3 Initial quality of the example transcription

Table 4-4 shows the percentage of symbols the ITs changed in the given example transcription. In order to get a feeling for the necessity for the ITs to change this percentage of symbols, we also measured the initial quality of the example transcription (ET). The quality of the ET, before the ITs modified this transcription, is established by comparing the ET with the consensus transcription. Table 4-5 shows the percentage of symbols in which the ET deviates from the consensus transcription.

Table 4-5 Percentages disagreement measured between the example transcription (ET) and the consensus transcription.

	RS	LC	IN	SC
ET	10.5	16.7	19.4	26.8

The proportion of symbols that needed to be changed in the ET is similar to the proportion of symbols that were actually changed in the ET by the four ITs (cf. Table 4-4). However, the inter-transcriber disagreement scores in Table 4-3 reveal that the ITs did not modify the

same symbols. It can then be concluded that the ITs changed symbols in the ET that were actually correct and left others intact that should have been modified.

4.4.4 Disagreement between individual and consensus transcriptions

The results presented in the previous section have shown the proportion of symbols the ITs changed in the ET and the level of disagreement between the individual transcribers in various speech styles. However, the figures do not reveal to what extent the transcriptions indeed reflect the speech signal; therefore we compared each IT with the consensus transcription. Four comparisons were carried out per speech style. The results are displayed in Figure 4-1 to Figure 4-4 for RS, LC, IN and SC, respectively. In each figure percentage disagreement is broken down into percentage of substitutions, deletions and insertions per individual transcriber. The percentage is calculated relative to the number of phonemes in the consensus transcription, which is of course the same in all four comparisons per speech style. In the right-handed panels the total percentages disagreement for each IT is given numerically. Extra columns are added for the average percentage disagreement (the hatched bars). In the bars that express the total percentage disagreement the two-sided 95% confidence interval is displayed.

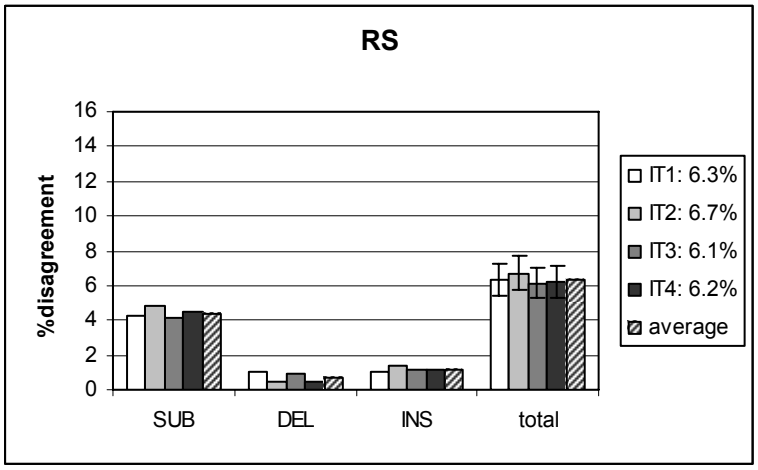

Figure 4-1 Percentages disagreement for read speech fragments between the individual transcriptions and the reference transcription. The hatched bars are the average percentage disagreement.

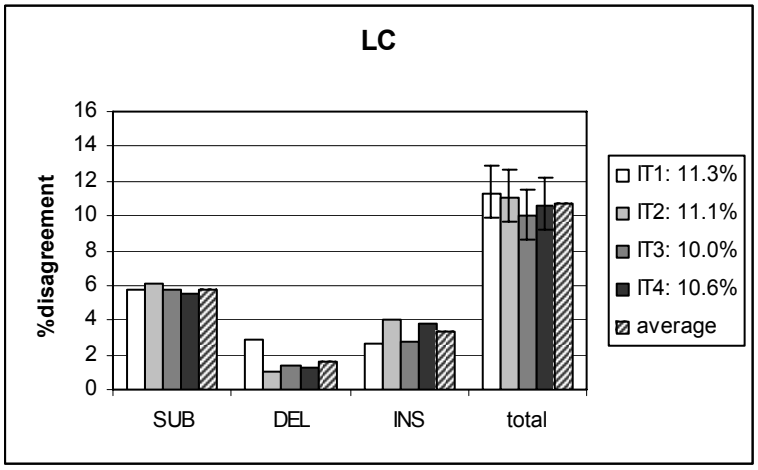

Figure 4-2 Percentages disagreement for lecture fragments between the individual transcriptions and the reference transcription. The hatched bars are the average percentage disagreement.

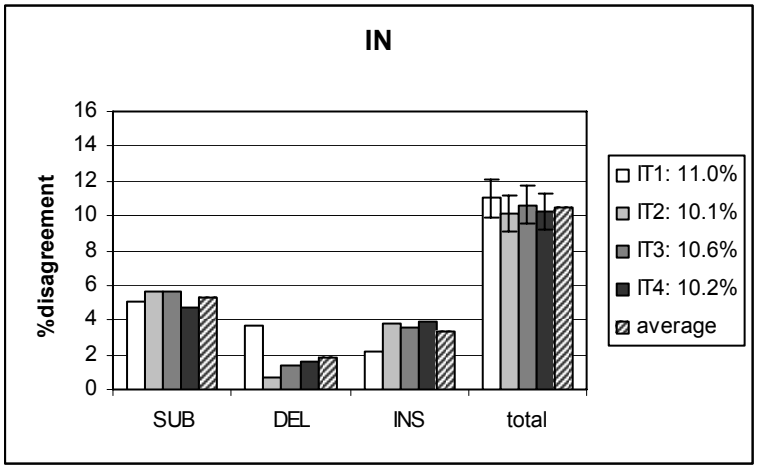

Figure 4-3 Percentages disagreement for interview fragments between individual transcriptions and the reference transcription. The hatched bars are the average percentage disagreement.

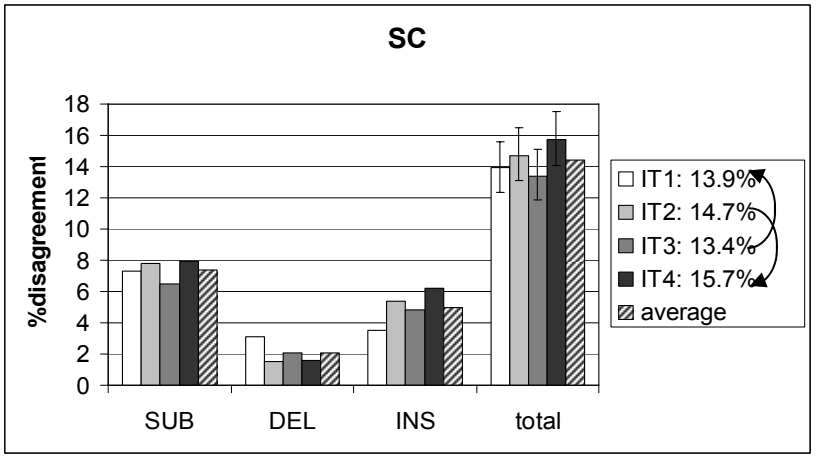

Figure 4-4 Percentages disagreement for spontaneous speech fragments between the individual transcriptions and the reference transcription. The hatched bars are the average percentage disagreement.

The results displayed in Figure 4-1 to Figure 4-4, clearly indicate that the distance between the ITs and the consensus transcription increases as the speech styles are characterised as more spontaneous. The ranges of 95% confidence intervals all overlap and therefore there are no significant differences between the percentages disagreement of the ITs in the four speech styles. The double-check procedure that was applied on the transcriptions of the spontaneous conversations (SC) made by IT1 and IT4 clearly solved some discrepancies. However, the percentages disagreement of IT3 and IT2 are within the one-sided 95% confidence intervals of IT1 and IT4, respectively, which implies that there are no significant differences between percentages disagreement before and after the double-check. On the other hand, for all speech styles it holds that the percentages disagreement between the four ITs and the reference transcription do not lie in the range of the one-sided 95% confidence intervals of the percentages disagreement between the ET (cf. Table 4-5) and the reference transcription. This implies that the transcribers improve the ET significantly in all four speech styles.

4.4.5 Articulatory distance between the individual transcriptions and the consensus transcription

It is impossible to determine from an agreement score computed on the number of symbols that differ between two strings, whether large or just minor articulatory differences were involved. Knowledge about the nature of the differences is instructive; minor articulatory differences imply that the transcription is closer to what was actually realised. The program *Align* uses articulatory information to make an optimal (lowest costs) alignment between two phoneme strings. Appendix A1 and A2 are matrices representing the articulatory

information, or costs, for consonsants and vowels, respectively. These costs stand for the seriousness of the discrepancy. For instance, substitutions between phonemes that share all but one articulatory feature (e.g. /d/ and /t/; difference in feature 'voice') have lower costs than substitutions between phonemes that differ in more features (e.g. /d/ and /v/; difference in features 'place' and 'manner' (stop and fricative)). In Figure 4-5 the percentage disagreement is plotted against the average articulatory costs per symbol that deviated between the consensus transcription and the ITs' transcriptions. The average articulatory cost displays the seriousness of the error.

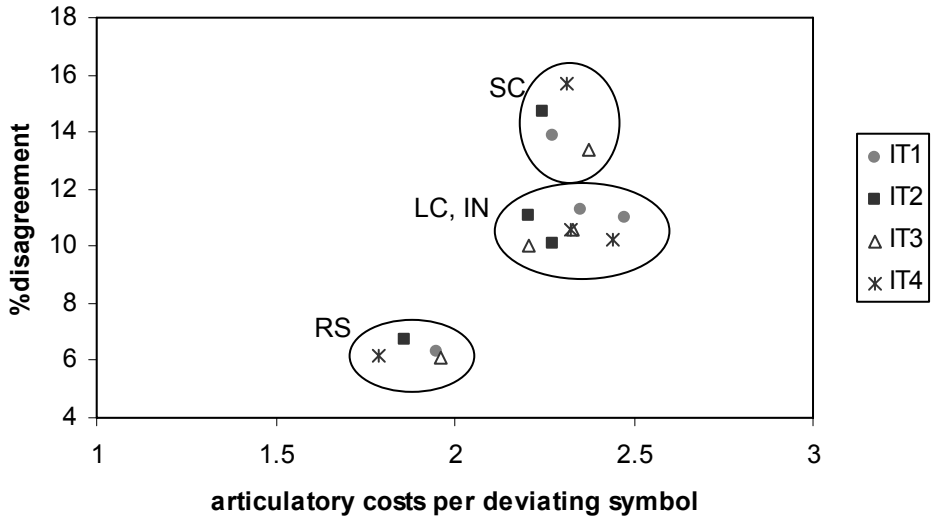

Figure 4-5 Percentage disagreement and articulatory costs per deviating symbol of all ITs' transcriptions in all four speech styles.

Figure 4-5 shows that within a speech style the seriousness, or the articulatory costs, per deviating symbol differs among the ITs. For instance, IT3 and IT4 deviate to a similar extent from the reference transcription in RS regarding percentage disagreement, 6.2% and 6.1%, respectively, but the transcription of IT4 is articulatorily closer to the reference transcription than IT3's transcription, given the lower cost per deviating symbol in IT4's transcription. More prominent in Figure 4-5 is the distinction between RS and the other speech styles. There is no clear distinction in seriousness of the errors in LC, IN, and SC fragments. The only distinction that can be made for these speech styles lies in the number of errors (percentage disagreement) that were made by the ITs. In other words, in transcribing more spontaneous speech, transcribers make more errors, but these errors are of the same order of seriousness. However, for RS fragments, the ITs not only make fewer errors (lower percentage disagreement) but these errors are also less serious compared to the other three speech styles. This means that the discrepancies between the ITs'

transcriptions and the reference involved fewer articulatory features. Although the articulatory costs already refine disagreement scores up to a certain extent, a qualitative analysis gives better insight into the type of articulatory differences, i.e. the nature of the errors.

4.5 Qualitative results

4.5.1 Inter-transcriber differences

Besides the number of discrepancies, the program *Align* also provides information about the nature of the discrepancies, revealing the type of substitutions, deletions and insertions. Table 4-6 presents only the five most frequent substitutions per speech style, which constitute over 50% of all substitutions in RS and well over 30% of all substitutions in the other speech styles. The symbols '[]' stands for unintelligible speech and are used when a segment could not be transcribed with sufficient confidence by that specific IT. In Table 4-6 a /x-G/-substitution of transcriber pair IT1-IT2 indicates that IT1 chose a voiceless /x/, where IT2 decided to transcribe a voiced /G/.

All ITs clearly had problems in deciding whether a velar fricative is unvoiced, /x/, or voiced, /G/. The problem did not appear in the SC fragments, which actually contained fewer instances of words that contain a velar fricative. It is remarkable that IT1 most disagrees with the other ITs with respect to this phenomenon, where IT1 chose the unvoiced velar fricative over the voiced variant. Whether or not IT1 made the right decision will become clear after the qualitative analysis of the differences between each of the ITs and the reference transcription in the next section. Still, it seems that IT1 followed another strategy in determining the voicing of the velar fricative. In the transcription protocol it is stated that a symbol in the example transcription should only be modified in case the ITs were absolutely confident that this symbol did not represent the speech signal adequately. Van de Velde (1996, p. 102-111) found some six different variants of the velar fricative in Dutch, while the ITs were forced to map these different variants on two variants that only differed in the feature voice. Whether or not the ITs followed the example transcription and trusted the phonological rules for applying /x/ or /G/, or whether they could distinguish the voiced and unvoiced variants remains unclear.

In general, the substitutions found between the ITs are quite similar within a speech style. Most of the frequent substitutions are related to a difference in the feature voice. For RS this type of substitution, in which two phones only differ in one feature, is responsible for all frequent deviations. In the other speech fragments other sorts of substitutions are also observed, such as confusion between full vowels and a schwa, or between long and

short vowels. The frequencies of these latter two types of substitutions are considerably higher in SC fragments.

Table 4-6 Five most frequent substitutions found between the individual transcribers in all four speech styles.

	IT1-IT2	IT1-IT3	IT1-IT4	IT2-IT3	IT2-IT4	IT3-IT4
RS	x-G	x-G	x-G	v-f	v-f	f-v
	f-v	z-s	f-v	z-s	z-s	s-z
	t-d	t-d	z-s	d-t	x-G	x-G
	a-A	v-f	t-d	A-a	f-v	t-d
	z-s	p-b	a-A	G-x	d-t	k-g
LC	x-G	x-G	x-G	v-f	v-f	o-O
	t-d	d-t	z-s	d-t	z-s	d-t
	f-v	z-s	d-t	z-s	d-t	a-A
	d-t	t-d	a-A	n-N	o-O	x-G
	N-n	k-g	o-O	E-@	n-N	z-s
IN	x-G	x-G	x-G	z-s	o-O	x-G
	t-d	@-A	o-O	v-f	x-G	o-O
	@-A	t-d	t-d	g-k	G-x	G-x
	f-v	o-O	@-A	@-A	z-s	A-@
	s-z	z-s	z-s	[]-n	[]-n	[]-m
SC	d-t	d-t	x-G	G-x	[]-n	x-G
	f-v	f-v	f-v	v-f	z-s	f-v
	a-A	a-A	a-A	d-t	d-t	k-g
	t-d	a-@	d-t	g-k	[]-m	t-d
	d-t	@-[]	k-g	A-a	[]-j	E-@

Since the ITs were compared pair-wise, the direction of insertions or deletions becomes meaningless and therefore we treated insertions and deletions as one type of discrepancy. This discrepancy type expresses the extent to which two transcribers disagreed upon the presence of a phoneme in the speech signal. The three most deleted and inserted phonemes were the same for all transcriber pairs within a speech style. Therefore, the data in Table 4-7 is not specified per transcriber pair, but per speech style. On average, the phonemes listed in this table comprise over 60% of all deletions and insertions made by each of the transcriber pairs. Since the chance of a phoneme to be deleted or inserted depends primarily on the absolute frequency of that specific phoneme in the sample, the relative

percentages of deletions and insertions are calculated. These relative percentages disagreement represent the percentages of occurrences of that specific phoneme that was either deleted or inserted.

Table 4-7 Three most deleted and inserted phonemes per speech style between two individual transcribers. The percentages express the relative number of deletions and insertion averaged over all transcriber pairs calculated on the number of occurrences of the phoneme in the consensus transcription.

RS		LC		IN		SC	
symbol	%	symbol	%	symbol	%	symbol	%
/r/	7.6	/r/	18.1	/r/	25.9	[]	72.6
/n/	2.0	/d/	8.4	/@/	11.2	/r/	21.8
/@/	1.3	/@/	6.3	/d/	9.1	/@/	14.1

Table 4-7 shows that the ITs differed in their judgements upon the presence (or absence) of the /r/, and /@/. Most of the /r/ disagreements are found in post-vocalic position at word ends (e.g. in the word 'voor') and word internally at prefix ends (e.g. in the word 'verschil'). Deciding whether a full /r/ was produced or whether the preceding vowel was r-coloured appeared to be difficult. With respect to the /@/, most disagreement was found in frequent short function words ('de', 'ze', and 'en'), at word ends, and in unstressed word internal positions ('gewoon'). Both /r/-deletion and /@/-deletion are well known phonological processes in Dutch (Booij, 1995; Van den Heuvel & Cucchiarini, 2001). A more remarkable phenomenon is the disagreement with respect to /d/. It often appeared that one IT decided that a word final /t/ was assimilated to /d/, while the other IT decided that the original /t/ was completely absent in the signal. Besides this, /d/-deletion, or insertion, also often occurred in very frequent short words ('de', 'dat', and 'd`r').

4.5.2 Differences between the individual transcriptions and the consensus transcription

From Figure 4-1 to Figure 4-4 it is clear that the majority of the differences between the individual transcriptions and the consensus transcription are due to substitutions and insertions, and to a lesser extent to deletions. The five most frequent substitutions made by each of the ITs are displayed in Table 4-8, covering about 35% of all substitutions made in each of the four speech styles. To clarify, a /v-f/-substitution means that the IT substituted the /v/ (as found in the consensus transcription) with an /f/. Most disagreement between the ITs and the consensus concern differences with respect to the feature 'voice'. This in line

with the finding that disagreements between ITs are mainly related to the feature 'voiced' (cf. Table 4-6). For the RS fragments all frequent substitutions involve voice substitutions, in which only one articulatory feature is involved. This explains the clear distinction between RS and the three other speech styles in Figure 4-5. The ITs tend to prefer the voiced variants of plosives (/d/ vs. /t/) and fricatives (/G/ vs. /x/) over the unvoiced variants in the consensus transcription.

Table 4-8 Five most frequent substitutions made by the individual transcribers opposed to the consensus transcription in four speech styles.

	IT1	IT2	IT3	IT4
RS	v-f	x-G	x-G	x-G
	s-z	t-d	t-d	t-d
	k-g	f-v	v-f	f-v
	t-d	s-z	z-s	k-g
	G-x	k-g	E-@	v-f
LC	s-z	x-G	x-G	x-G
	t-d	t-d	t-d	k-g
	k-g	k-g	k-g	t-d
	O-o	s-z	@-E	@-E
	A-a	@-E	O-o	o-O
IN	k-g	t-d	t-d	t-d
	t-d	x-G	x-G	x-G
	O-o	@-A	k-g	k-g
	@-A	k-g	@-A	@-A
	I-@	I-@	I-@	I-@
SC	t-d	x-G	@-E	k-g
	@-E	k-g	k-g	x-G
	k-g	t-d	t-d	t-d
	A-a	@-E	@-A	@-E
	v-f	s-z	A-@	@-A

The inventories of substitution types in both Table 4-6 (inter-transcriber substitutions) and Table 4-8 are quite similar; however, the ranking of the substitutions differs, which is an indication that the ITs each concentrated on a limited, but potentially different number of phenomena. We can illustrate this by singling out the substitutions of /x/ and /G/. As mentioned above, compared to all other ITs in Table 4-6, IT1 preferred the /x/ over the /G/.

In Table 4-8 we can see that IT1 indeed appeared to have made the right choice in this respect, and that the 'preference for /G/' of the other ITs led to more deviations from the consensus transcription (many /x-G/ substitutions). IT1 seems to have followed the same strategy as the expert transcribers in the consensus transcription in distinguishing the /x/ from the /G/.

Table 4-9 Absolute number of /t-d/ and /x-G/ substitutions for ITs and ET.

	REF	ITs & ET	IT1	IT2	IT3	IT4	ET
RS	t	d	8	16	14	17	11
	x	G	3	26	23	27	23
LC	t	d	16	22	19	15	14
	x	G	3	26	24	23	25
IN	t	d	7	11	10	10	10
	x	G	1	11	9	9	8
SC	t	d	16	15	10	13	9
	x	G	2	17	4	15	9

The two most frequent substitutions occurring in all four speech styles with respect to all four ITs (Table 4-8) are /t-d/ and /x-G/. The prominence of these substitutions called for a closer examination in order to learn more about the transcription strategies of the ITs. We calculated the number of these substitutions occurring between the example and the consensus transcription in order to see if there is any similarity in this respect between the ITs and the example transcription. The second column in Table 4-9 shows the label found in the consensus transcription (REF); the third column gives the label as occurring in the ITs and the ET. The following columns show the absolute frequency of these labels in all four ITs and the ET. In some cases the IT corrected the 'errors' in the example transcription, e.g. IT1 changed /G/ to /x/, according to the consensus transcription. However, in most other cases the ITs kept the variant that was present in the ET (in this example /G/ and /d/), which proved to be different from the consensus transcription. Moreover, when looking at the actual transcriptions (phoneme strings) it appeared that the errors of the ITs originate from leaving the ET intact at almost the same words, and from incorrectly changing the ET at other points, which explains the higher number of the /t-d/ and /x-G/ substitutions of the ITs compared to the ET, at least for IT2, IT3, and IT4.

The biasing effect of the ET that is reflected in the high similarity between the ITs and the ET can be caused by several factors. We believe that the ITs simply cannot focus on all

processes (a large number, especially in the non-read speech styles), because other phenomena could have been, in their view, more important or more salient. After all, the ITs were expected to work at a considerable speed. At the same time, we can assume that sometimes the ITs were incapable of deciding which symbols (voiced or unvoiced) best matched the speech sound, and for that reason left the ET intact. Similar to /x/ and /G/, the acoustic difference between /t/ and /d/ is not only the result of differences in vocal fold vibration alone. Van den Berg (1988) and Slis (1985) found eleven acoustic differences between /t/ and /d/ of which only two pertain to characteristics of the part of the signal that directly corresponds to the /t/ and /d/, of which vocal cord vibration was one. In these cases different transcribers can very well use different strategies and criteria for mapping a complex signal onto a binary choice. It is plausible that both IT1 and the expert transcribers in the consensus transcription used similar criteria, while IT2, IT3, and IT4 used other procedures and criteria for making their decision.

The other source of discrepancies between the ITs and the consensus transcription were the insertions. Insertions mean that the ITs used more symbols in their transcription than were present in the consensus transcription. A closer look at the insertions in all speech styles reveals that all ITs seemed to have problems deciding on three phonemes, viz. /@/, /r/, and /n/, which also came to light in the qualitative analysis of the inter-transcriber disagreement figures. The relative contribution of these insertions to the total number of insertions per IT is on average about 62% for all four speech styles. Examining the ET in this respect, we found that four phonemes were frequently 'inserted', viz. /r/, /@/, t/, and /n/. It seems as if the ITs have solved the /t/-insertion problem to a certain extent. Relating this finding to the prominent disagreement on the presence or absence of the /d/ found between the ITs, see Table 4-7, we can make the assumption that the ITs focussed more on /t/-deletion, or assimilated /t/-deletion (/d/-deletion), than on /n/-deletion, a process that occurred more often than the ITs seem to have noticed. Word-final /n/-deletion after /@/ is a common phonological process in Dutch. In the canonical forms used for the generation of the example transcription, this phonological process was not implemented. In this respect the ITs again kept the example transcription intact, a tendency also observed in Demuynck et al. (2004) on the exact same phenomenon for Dutch as spoken in Flanders.

The results of the qualitative analysis presented above indicate that agreement indices based on the number of symbols that differ between two strings do not provide a complete measure of transcription quality. These global measures, expressed in percentages, do not present the full picture. At least for some applications of phonetic transcriptions (for example a study into gender related voicing of plosives and fricatives) the percentage agreement is likely to suggest a substantially higher quality than is actually available.

4.6 Discussion

In processing large amounts of speech data that need to be enriched with manual phonetic transcriptions it is common practice to adopt a transcription procedure in which transcribers make an auditory transcription by editing a given example transcription. More often than not, different parts of the corpus are transcribed by different transcribers (often students). The main objective of this study was to evaluate this procedure by analysing an instantiation of it, namely the transcriptions produced in the framework of the CGN project. Before any judgements about the quality of the transcription can be made, one must first decide how to measure transcription quality. For this purpose two methods were adopted: a) measuring inter-transcriber agreement, which is a widely used measure of transcription quality in the context of large speech corpora, and b) determining the distance between the individual transcriptions and an approximation of a true reference transcription. In this study we also argued for additional qualitative analyses of the differences, so as to go beyond the summary nature of agreement scores.

In the literature inter-transcriber agreement is often used as a measure to assess transcriptions (Eisen, 1993; Greenberg et al., 1996; Kikuchi and Maekawa, 2003; Pitt et al., 2005; Raymond et al., 2002; Wesenick and Kipp, 1996). Our results on inter-transcriber (dis)agreement are in line with agreement scores reported in those papers. Eisen (1993) found 90% inter-transcriber agreement for consonants and about 83% for vowels in read speech (German), where transcriptions were produced by editing an example transcription on a broad phonetic level. Greenberg et al. (1998) report 75%-80% inter-transcriber agreement for American-English spontaneous telephone conversations. Again, the transcriptions were produced by verifying an example transcription, but a narrower symbol set was used, the Arpabet symbol set. Raymond et al. (2002) and Pitt et al. (2005) both examined human transcriptions (made with the DARPA phonetic alphabet) in the Buckeye corpus (American-English spontaneous dialogues) and report agreement scores of 76%-80.3%. And finally Wesenick & Kipp (1996) found 94.8% agreement for consonants (89.9% for stops, 98.0% for fricatives and 97.5% for nasals) in read speech (German) by cross validating the individual transcriptions. The tasks for the transcribers were different (other phoneme set, different languages, add time-alignments) in the above studies, but they all had to verify an example transcription. In Coussé et al. (2004) student transcribers had to transcribe Dutch vowels in spontaneous speech fragments of the CGN without having an example transcription at their disposal. Inter-transcriber agreements indices of 49.1% to 59.6% were reported. The speech material is similar to the material used in our experiment, viz. the spontaneous speech (SC) fragments. However, the ITs in our

experiment, who were student transcribers from the CGN project as well, had to correct a given example transcription on all possible phonemes. Since the task was not the same, the figures reported by Coussé et al. (2004) are in contrast with the inter-transcriber agreement percentages we reported in this paper (4.7% to 12.5% on average on read speech and spontaneous speech, respectively). No experiments were conducted to measure and to quantify the biasing effect of the given example transcription, but it is clear that we can assume that the inter-transcriber agreement percentages we reported in this paper are somewhat inflated by the presence of the example transcription.

The main objective of this study was to determine whether inter-transcriber agreement is an appropriate measure to establish the quality of transcriptions made by verifying and correcting an example transcription. Inter-transcriber agreement expresses the degree of similarity in phonetic transcriptions of the same speech sample produced by different transcribers, in other words, it expresses the consistency of the transcriptions. There are two important factors that determine expected agreement (cf., Shriberg & Lof, 1991). Firstly, the degree of detail in transcription symbols; narrow versus broad transcriptions, and secondly, the bias of the given example transcription.

In this study we used a consensus transcription as an additional measure for assessing transcription quality. The consensus transcription was made by two transcribers who agreed on each single symbol in the transcript, which resulted in a minimisation of transcribers' subjectivity and idiosyncrasies. Owing to this, the resulting consensus transcription approaches the ground truth. Moreover, the minimisation of idiosyncrasies and other subjective judgements, allows us to consider the consensus as some kind of average transcription. Our results show that the ITs do indeed improve the example transcription, since the deviations between the consensus transcription and the example transcription (initial quality) are significantly larger than those between the consensus transcription and the ITs. However, the average percentages of deviations between the ITs (inter-transcriber disagreement) in RS, LC, and SC are significantly smaller than the average distance between the ITs and the consensus transcription, despite the fact that individual transcriptions may suffer from subjectivity and idiosyncrasies. This provides further evidence for the fact that the given example transcription had a biasing effect on the transcribers' judgements, such that the ITs show higher agreement scores than expected.

Some trends emerge from our more detailed analyses, in which we examined the articulatory distance between transcriptions and the nature of the individual discrepancies. For instance, we found that read speech is in general easier to transcribe, not only because fewer mistakes are made, but also because the mistakes are less serious than those made in

more spontaneous speech styles. In read speech the pronunciation of the words resembles the canonical pronunciation, which was the basis for the example transcription, more than any of the other speech styles. In addition, qualitative analyses of the differences found among the individual transcriptions and between the individual transcriptions and the consensus transcription reveal that the most salient type of substitution is related to the feature voice; the individual transcribers frequently disagree with each other as well as with the consensus transcription. This suggests that student transcribers find it difficult to distinguish between voiced and voiceless consonants. This reflects the uncertain status of the Dutch fricatives with respect to the feature 'voice' (Gussenhoven, 1981; Kissine et al., 2003). For the plosives it holds that there are a large number of acoustic phonetic features involved in the voiced-unvoiced distinction (Van den Berg, 1988; Slis, 1985) which are likely to add to the disagreement between independently working transcribers. In order to reach a high level of agreement, it is necessary to develop a common frame of reference during a training phase. In the CGN project there was little room for such a training phase (although the ITs went through a short training supervised by one of the persons who were involved in creating the consensus transcription).

A second source of discrepancies between the individual transcriptions is the decision whether or not /@/, /r/, /d/ and /n/ are present. These phonemes are difficult to detect since their presence is often encoded in potentially subtle changes in surrounding phonemes (r-colouration) or because they are extremely short. Again, high levels of agreement can only be reached at the cost of extensive training before the transcriptions are started. Finally, as in Demuynck et al. (2004), we found that the student transcribers left the example transcription intact too often. This can be directly related to the nature of the assignment to the transcribers, i.e., only to change the example transcription when the transcriber was confident that it was not properly representing the actual speech signal. Furthermore, the transcribers were expected to work under certain time pressure.

The additional analyses presented in this paper, the phonetic details of the differences, show that the measure 'percentage agreement' between two transcriptions should be treated with caution, in the sense that it is a global figure that does not represent the underlying processes (that potentially could change existing impressions on transcription quality).

To recapitulate, transcriptions produced according to the method followed in the CGN, and in many other large speech corpora, are indeed closer to a reference transcription than to the initial (example) transcription, but the bias of the example transcription should not be underestimated. Moreover, despite the money-saving procedure followed in the CGN, recourse to human transcribers even if it is limited to editing an example transcription, is still time-consuming and expensive whereas its added value is not always clear. Van Bael

et al. (2005), for example, showed that human-made transcriptions in the CGN for read speech yielded similar word error rates in an automatic speech recognition system as automatically generated transcriptions of the same material. Especially in the field of automatic speech recognition the amount of data is much more important than marginal improvements (if any) of the transcription quality made by a human transcriber, which consequently puts restrictions on the amount of data that can be processed in a reasonable time span.

4.7 Conclusions

Based on the above discussion of the results, the following conclusions can be drawn. First, the decision to employ human transcribers to correct an example transcription should be dependent on the type of speech that needs to be enriched with phonetic transcriptions. For instance, for spontaneous conversational speech the added value of human transcribers in the transcription generation process proved to be larger than in speech styles containing less pronunciation variation, such as read speech. Although human transcribers significantly improved the example transcription of read speech, given the time and costs, and the fact that these student transcribers may use different criteria to distinguish between voiced and unvoiced phonemes, we suggest that future projects on transcriptions of large amounts of speech data should think twice before hiring expensive human transcribers for phonetic transcriptions of read speech.

Second, the double-check procedure as was applied for transcriptions of spontaneous conversations in the CGN project, improved transcription quality. However, since this improvement was not significant, we suggest omitting this expensive procedure in future projects.

Third, if transcriptions are made by editing example transcriptions, human transcriptions, of any type or quantity, should not be evaluated by solely establishing an inter-transcriber agreement score score, since the effect of the example transcription is likely to be so strong that the ITs cannot be considered as independent observations. Rather, additional measures, such as a comparison with a consensus transcription, should be taken into account as well.

Acknowledgements

We would like to thank the two expert phoneticians and the four CGN transcribers for their contribution to this study.

MULTIWORD EXPRESSIONS IN SPOKEN LANGUAGE

CHAPTER 5

Reformatted from:

Diana Binnenpoorte, Catia Cucchiarini, Lou Boves and Helmer Strik (2005). Multiword expressions in spoken language: an exploratory study on pronunciation variation. *Computer Speech and Language 19(4),* pp. 433-449.

The study presented in this paper was aimed at exploring the possibilities of modelling specific pronunciation characteristics of multiword expressions (MWEs) for both automatic speech recognition (ASR) and automatic phonetic transcription (APT). For this purpose we first drew up an inventory of frequently found N-grams extracted from orthographic transcriptions of spontaneous speech contained in a large corpus of spoken Dutch. These N-grams were filtered and subsequently assigned to linguistic categories. For a small selection of these N-grams we examined the phonetic transcriptions contained in the corpus. We found that the pronunciation of these N-grams differed to a large extent from the canonical form. In order to determine whether this is a general characteristic of spontaneous speech or rather the effect of the specific status of these N-grams, we analysed the pronunciations of the individual words composing the N-grams in two context conditions: 1) in the N-gram context and 2) in any other context. We found that words in N-grams do indeed have peculiar pronunciation patterns. This seems to suggest that the N-grams investigated may be considered as MWEs that should be treated as lexical entries in the pronunciation lexicons used in ASR and APT, with their own specific pronunciation variants.

5.1 Introduction

Multiword expressions (MWEs) have been studied in theoretical linguistics (Nunberg et al., 1994; Sag et al., 2001; Wong-Fillmore, 1979), and more recently also in NLP (Koster, 2004; Nivre & Nilsson, 2004; Odijk, 2004). So far, most of the research on MWEs has concerned their extraction and handling in written language. However, it has also long been known that frequently used sequences of words, whether they are stock phrases (e.g. 'I don't know') or lexicalized idiomatic expressions (e.g. 'kick the bucket'), show pronunciation phenomena that have not been observed when the words occur in less frequent contexts (cf. the pronunciations of 'I don't know' in Hawkins (2003)). While observations such as Hawkins' are to some extent anecdotal, the advent of large spoken language corpora has made it possible to investigate pronunciation variation in multiword expressions quantitatively.

In this paper we investigate pronunciation variation in MWEs in a large corpus of spontaneously spoken Dutch (Oostdijk, 2002). Although the Spoken Dutch Corpus (also known as CGN) also comprises more formal speech styles, we focus on spontaneous speech because we think that the problem of pronunciation variation in MWEs is most acute in this style. Speech recognition performance for spontaneous speech is way below the performance for read speech (Pallett, 2003) and there are indications that a large

proportion of the performance gap is due to the inability to model pronunciation variation in spontaneous speech effectively (Strik & Cucchiarini, 1999).

For ASR it has been found that simply adding the most frequent pronunciation variants of individual words to the lexicon becomes counter-productive as soon as the average number of variants per word exceeds a threshold of about 2.5 (Kessens et al. 2003; Yang & Martens, 2000b). At the same time, it appears that adding frequent bigrams to the lexicon and treating these as words with their own specific pronunciation variants does improve ASR performance (Beulen et al., 1998; Finke & Waibel, 1997; Kessens et al., 1999; Sloboda & Waibel, 1996). However, in these studies the notion of MWE is mainly deployed for the benefit of reducing word error rate in ASR. No special attention was given to the lexical and linguistic role and status of the word sequences. In the present paper we investigate whether it is indeed true that words in MWEs in spontaneous speech have more -and specifically more reduced- pronunciation variants than when the same words occur in a general context.

In our research we first extracted frequent word sequences (which we will call MWEs for convenience throughout this paper) from all spontaneous speech recordings in the Spoken Dutch Corpus (CGN), which we then analyzed to determine their lexical status and syntactic structures. Then we proceeded to a more detailed analysis of MWEs in that part of the CGN that comes with manually verified broad phonetic transcriptions. In doing so, we focused on reduction phenomena, and we tried to determine whether there is a relation between the degree of reduction in a given MWE and the lexical/syntactic category to which it belongs.

5.2 MWEs in the Spoken Dutch Corpus

MWEs were extracted from the Spoken Dutch Corpus, a database containing about 9 million words of contemporary Dutch as spoken in the Netherlands and Flanders. All recordings are orthographically transcribed, lemmatised and enriched with part-of-speech (POS) information. For about 900,000 words, more detailed annotations are available, such as a manual broad phonetic transcription, a hand-checked word alignment, syntactic annotation and prosodic information. This sub-corpus of 900,000 words, called the *core corpus*, was composed in such a way that it faithfully reflects the design of the full corpus (Oostdijk, 2002). The speech material in the corpus was recorded in various socio-situational settings from speakers of different age, sex, educational level and region of birth. The speech material collected consists of various speech styles, varying from read speech recorded in a studio environment with professional speakers, through interviews

which are more or less prepared dialogues, and business negotiations to spontaneous dialogues recorded in home environments.

For our study we are only interested in spontaneous speech; therefore, only speech styles that can be characterized as spontaneous or extemporaneous were selected. In order to make a comprehensive inventory of MWEs in unprepared speech, we used the orthographic transcriptions of all lessons (LS), spontaneous dialogues (SD), and spontaneous telephone conversations (ST). The conversational settings differ among the three components. In the LS component a teacher discusses and explains several subjects with a group of students. In the SD component two or more people have a face-to-face conversation in a home environment, often about objects in the room or activities such as game playing that they are involved in. Finally, in the ST component two friends or family members have a telephone conversation without the need to talk about specific topics. Table 5-1 summarizes the characteristics of the material, selected from the Nothern-Dutch part of the corpus, that are most important for the present study.

Table 5-1 Total duration of the components, number of words and number of different speakers involved.

speech style	duration (hh:mm:ss)	#words	#speakers
LS	30:41:04	299,973	398
SD	149:44:17	1,747,789	231
ST	92:24:50	1,253,741	534
total	272:50:11	3,301,503	1,148

5.2.1 Criteria for selecting N-grams as MWEs

There is no generally accepted definition of the concept of MWEs in spoken language. Therefore we based our investigations on what we consider a reasonable operational definition of the concept, adapted to the specific requirements of our study. Since we are interested in the effect of MWE status on pronunciation variation, our first criterion was that only contiguous sequences of words qualify. We expect to see substantial pronunciation variation in the form of cross-word assimilation and degemination. In lexicalized MWEs that are broken by interspersed words, the cross-word phonetic context of the contiguous MWE no longer exists. Consequently, one cannot expect to observe the cross-word assimilations and reductions that may be characteristic for the contiguous MWEs. A practical advantage of this criterion is that it allows us to start the search for

potential MWEs by simply creating lists of sequences of *N* words with a frequency of occurrence that is higher than what one would expect for arbitrary syntactically correct sequences.

Thus, we started the search for N-grams that might qualify as MWE by extracting all 3-4-, 5-, and 6-grams from the orthographic transcription files. In doing so, we used the – admittedly somewhat arbitrary – criterion proposed in chapter 13 in Biber et al. (1999) to establish the minimum frequency that a sequence should exceed in order to qualify as 'exceptionally frequent'. Expressions containing three or four words should have a minimal frequency of 10 per million words, and expressions containing more than four words should have 5 or more occurrences per million words. In our case, with a source text of 3.3M words, we require the frequency of a unique 3-gram and 4-gram to be at least 30, and for the 5-gram and 6-gram at least 15.

Because we want to use frequent sequences to investigate pronunciation variation in word sequences that may qualify as MWEs, or at least as stock phrases, we decided to apply a number of additional criteria to filter the raw lists of expressions that exceed Biber's frequency threshold. First, we did not want to include word sequences that straddle a deep syntactic boundary. These are likely to induce pauses between the words on either side of the boundary that block assimilation and degemination processes. The only clues for syntactic boundaries in the CGN transcriptions are full stops, question marks, and ellipsis marks; no commas and other 'minor' punctuation marks are included. Therefore, we restricted the search for MWEs to sequences that do not include one of the three punctuation marks.

A second criterion in the filter process was the length of the sequences. Given the size of the corpus, we did not expect to find frequent sequences longer than six words. For theoretical and practical reasons we decided to omit bigrams. For one thing, many frequent bigrams are part of frequent N-grams with $N > 2$, so that we can observe and analyze their pronunciation variation even if we do not include bigrams. Moreover, the number of frequent bigrams is extremely large, and the sheer number complicates analysis considerably. Therefore, we decided to take $3 \leq N \leq 6$.

Third, we decided to exclude disfluencies and hesitations from our corpus of frequent N-grams. The initial N-gram list contained a substantial number of frequent sequences in which one or more filled pause markers were present. In the CGN all filled pauses are transcribed by one of two 'hesitation' words, 'uh' and 'uhm'. This transcription convention is part of the explanation why word sequences containing filled pause markers occurred so frequently. Another part of the explanation is definitively related to the fact that filled pauses and hesitations do not occur in random positions, but tend to occur just before content words, due to which sequences such as 'in the ehr' are rather frequent. Although

detecting and handling hesitations and disfluencies is of crucial importance for automatic recognition of spontaneous speech, we feel that these phenomena form a research topic in their own right, probably related, but also somewhat independent of pronunciation variation in MWEs. Therefore, we excluded N-grams such as 'de uh de uhm' ('*the eh the ehr*') as potential MWEs. Sequences containing 'ggg' (the symbol for speaker noise) or 'xxx' (unintelligible speech) were excluded for the same reason.

Fourth, we also decided to exclude repetitions. In the spontaneous part of the CGN one can distinguish two different categories of repetitions. The first category, which comprises sequences such as 'en de en de' ('*and the and the*'), represents what are likely to be disfluencies. These cases are rejected for the reason explained above. The second category is perhaps more problematic. It contains sequences such as 'ja ja ja ja' ('*yes yes yes yes*'), which may be related to disfluencies, but which can also be used to indicate emphasis or other pragmatic effects. The CGN transcriptions do not provide information that can be used to distinguish disfluencies from truly linguistic devices, such as for lending emphasis or expressing sarcasm. For this reason we decided to remove all two and three word repetitions from the lists of possible MWEs.

The last criterion that we used to filter the lists of frequent N-grams is the requirement that the sequence should have higher than expected frequency in all three sub-corpora (LS, SD, ST). This stipulation removes sequences such as 'een twee drie vier' ('*one two three four*'), which are frequent in the SD sub-corpus, due to the fact that the speakers were encouraged to play games to keep the conversation going. Perhaps it might be possible to identify and eliminate setting-specific sequences on the basis of linguistically informed rules, but it is very difficult to formulate adequate rules. Thus, we used the uniform presence criterion to detect and remove such artefacts from the lists.

Table 5-2 summarizes the results of the MWE extraction on the 3.3M word spontaneous speech part of the CGN. It can be seen that both the number of types and the token/type ratio decrease as the sequences grow longer. The number of types would have been much larger if we had not applied the criterion that expressions should occur with higher than expected frequency in all three sub-corpora. That criterion removed many sequences from the sub-corpus of face-to-face dialogs that were directly related to playing card or board games. Removing setting specific types resulted in a large increase in the average token/type ratio.

From Table 5-2 it can be deduced that the 3,311 N-gram types cover about 21% of the source corpus. Apparently spontaneous conversations consist to a large extent of 'stock phrases' and/or true MWEs. As not many generalisations can be made over one type, the one remaining 6-gram will not be considered in the remainder of the paper.

Table 5-2 Number of types and tokens of N-grams passing the selection criteria.

	3-grams	4-grams	5-grams	6-grams
#types	3,015	247	48	1
#tokens	217,230	13,495	1,285	19
token/type ratio	72.05	54.63	26.71	19

5.2.2 Categorization of selected N-grams

Once the MWEs had been extracted from the transcription files, we proceeded to classify them manually into six broad categories:

1. The N-gram constitutes a whole grammatical sentence.
   e.g. 'weet ik veel' ('*I've no idea*')
2. The N-gram constitutes a grammatical constituent.
   e.g. 'op een andere manier' ('*in a different way*')
3. The N-gram constitutes an interjection.
   e.g. 'nou ja goed' ('*well alright*')
4. The N-gram constitutes the beginning of a possible main clause.
   e.g. 'en dan moet je' ('*and then you have to*')
5. The N-gram constitutes the beginning of a possible subordinate clause.
   e.g. 'als het goed is' ('*if it is okay*')
6. The N-gram cannot be classified in any of the above and is categorized as 'other'.
   e.g. 'weet niet of je' ('*don't know whether you*')

These categories emerged during the process, based on our interpretation of the MWEs. The categories fall apart in two broad classes; the first three categories include complete syntactic units, whereas the last three include sequences of words that do not constitute a complete syntactic unit. The distribution of the categories of the MWE types is displayed in Table 5-3.

Although the classification results in Table 5-3 are instructive, it should be noted that many MWEs assigned to the categories 2 to 5 would be moved to another class if some highly frequent function word were added before or after the sequence. Thus, the classification is to some extent based on evidence that is not extremely reliable. It would be worthwhile to repeat the experiment with a mix of words and POS information, and count

the frequency of sequences of the form *POSx, word1, ... wordn* and *word1, ... wordn, POSy*, where *POSx* indicates a set of words with the POS-tag *x*.

Table 5-3 Distribution of categories expressed in number and percentage.

	3-gram		4-gram		5-gram	
	#	%	#	%	#	%
1. complete sentence	163	5.4	25	10.1	9	18.7
2. constituent	260	8.6	18	7.3	3	6.3
3. interjection	64	2.1	12	4.9	5	10.4
4. begin of main clause	1002	33.2	124	50.2	22	45.8
5. begin of subordinate clause	126	4.2	4	1.6	0	0.0
6. other	1537	51.0	71	28.7	14	29.2
total	3152	104.5	254	102.8	53	110.4
categorized twice	137	4.5	7	2.8	5	10.4
# types	3015	100.0	247	100.0	48	100.0

Some trends emerge from this table. In general, for all three N-gram types, the contribution of N-grams classified as incomplete syntactic units (category 4, 5, and 6) is much larger than the contribution of those classified as complete syntactic units. During the selection procedure no restrictions on syntactic completeness were applied, because syntax annotation is only available for the *core corpus* in the CGN. Moreover, in Kessens et al. (1999) it is shown that modelling pronunciation variation of highly frequent sequences of words does improve recognition performance, but these word sequences need not constitute syntactic units.

The majority of the N-grams belong to category 4, where the N-gram constitutes the beginning of what is likely to become a main clause. In Dutch given information tends to go to the beginning of a clause, whereas new information tends to occur at the end. The high proportion of conventional expressions at the beginning of a clause may well help speakers to overlap cognitive processing needed to express the new information with almost automatic generation of the beginning of the sentence or clause in which the new information is embedded. Listeners may also profit from such an alternation of predictable and new information. In any case, the high frequency of a small number of clause-initial 'formulae' suggests that in conversational Dutch the variety of introductory clauses is not very broad. This impression is corroborated by the fact that the average number of tokens per type in the N-grams in category 4 is relatively high. Therefore, the frequently used N-

grams at the start of a main clause actually occur more often than might appear from the figures in Table 5-3, which only refer to types.

In the collection of the 3-grams the proportion of the 'other' category is larger than that of 'begin of main clause'. This might indicate that a sequence length of three words is too short to be identified as a possible beginning of a main clause or even a syntactic unit. This hypothesis is in line with the observation that adding one word at the beginning or end of a sequence often would change its category assignment. Alternatively, conversational speech may contain a substantial number of frequent word sequences that straddle the boundary between NP, PP or AP[I] constituents. Future research, in which POS (and perhaps also syntactic annotation) is used will show which possibility is more likely.

When the length of the sequences increases, the share of complete sentences and multiword interjections (category 1 and 3) also increases. The prominent presence of long interjections motivated the creation of category 3, as a special case of category 2 during the course of the classification process. In this context it is interesting to observe that the proportion of complete grammatical constituents which are not a sentence or an interjection decreases when the sequence length grows. This may indicate that highly frequent constituents (NPs, PPs and APs) mainly consist of three words in conversational Dutch.

5.3 Pronunciation variation in MWEs

Having compiled the lists of MWEs and some data on the occurrences extracted from the spontaneous speech in the CGN, we proceed to investigate whether words in MWEs have more reduced pronunciation variants than when the same words occur in another arbitrary context. This part of the study is limited by necessity to the Northern Dutch part of the *core corpus* in CGN, i.e., the part that comes with manually verified broad phonetic transcriptions. On average, the *core corpus* covers 10% of the total corpus. In Table 5-4 the size and other characteristics of the spontaneous components of the *core corpus* are displayed. From a comparison with the figures in Table 5-1 it can be seen that the spontaneous speech styles are represented more or less proportionally in the *core corpus*.

[I] noun phrase, prepositional phrase, and adjective phrase respectively

Table 5-4 Duration, number of words and number of different speakers in the spontaneous components of the core corpus.

speech style	duration (hh:mm:ss)	#words	#speakers
LS	2:43:36	25,961	48
SD	9:43:39	106,182	108
ST	14:42:28	201,141	101
total	27:09:43	333,284	255

5.3.1 Selection of frequent N-grams for pronunciation analysis

The analysis of the effect of the frequency of N-grams on pronunciation variation can only be performed on those N-grams that occur sufficiently frequently to allow us to distinguish systematic from coincidental observations. This issue is all the more urgent since we now must work with a corpus of no more than 0.3 M words. There are no formal criteria to determine what 'sufficiently frequent for the purpose of analyzing pronunciation variation' is. However, it is clear that we need an absolute lower bound, in addition to the relative lower bound proposed in Biber et al. (1999) for other types of linguistic analyses. To start the analysis we decided to restrict our corpus to types which occur at least 7 times. We considered this as the minimum number that should allow at least some conclusions about the characteristics of pronunciation variants. In the 0.3M word corpus of manually transcribed spontaneous speech there were no 5- or 6-grams that fulfilled this minimum frequency criterion. Consequently, the remainder of this paper is limited to an analysis of 3-grams and 4-grams. In Table 5-5 the number of different N-grams for which at least 7 observations were found is displayed for the 3-grams and 4- grams, together with the mean frequency and the frequency range.

Table 5-5 Properties of remaining N-grams.

	3-gram	4-gram
# types	110	21
mean frequency	17.5	13.8
frequency range	7 – 118	7 – 50

We can now proceed to making an inventory of the pronunciation variants of the words that occur in frequent N-grams. The *core corpus* provides word segmentations, which

connect the speech to the orthographic and phonetic transcription on the word level. This allows us to determine an unambiguous phonetic transcription for each word in the orthographic transcription.

5.3.2 Method of pronunciation analysis

Before we can proceed to the results of our analysis of pronunciation variation, we must first deal with two further methodological issues, viz. the way in which we defined the reference material to which we compared the pronunciation variants observed in frequent N-grams and the measure used to express differences in pronunciation variation.

Selection of reference material

To determine whether words occurring in frequent N-grams indeed have pronunciation variants that are different from the variants that can be observed for the same word in arbitrary but comparable contexts, we have to define the very concept *arbitrary but comparable context*. Ideally, one would like to compare words in the same syntactic and prosodic context, only now surrounded by other words that do not form a frequent N-gram. However, since the CGN *core corpus* does not provide sufficient prosodic and syntactic information, we decided to settle for a less ambitious definition. For each word we performed an N-gram search with the restriction that only N-grams were allowed in which that specific word was in exactly the same position as in the original N-gram and that the other words in the N-gram were different from those in the original N-gram. For instance, assuming that the word 'als' as found in the 3-gram 'als het ware' ('*as it were*') is subject to this detailed analysis (because the 3-gram 'als het ware' is one of the highly frequent N-grams) then only those versions of 'als' are taken into consideration in which the two words following 'als' do not equal 'het' and 'ware'.

Comparing different transcriptions

In order to compare the degree of discrepancy found in the conditions 'only within MWE context' and 'in all other contexts' (indicated as 'MWE context' and 'other context', respectively, in the remainder of the paper) we used the canonical transcription of each word as a reference point. More specifically, we compared the transcription of the words in the N-gram context to their canonical transcription, and we did the same with the occurrences of the words in arbitrary contexts. In this way we were able to calculate the weighted average percentage of difference for each word in the two conditions, where the weighting is based on the length of the word in question (number of segments in canonical transcription).

The differences between actually observed pronunciations and canonical representations were determined by the computer program *Align* (Cucchiarini, 1996). Table 5-6 shows the orthographic and canonical phonemic representations of the 4-gram 'aan de andere kant' ('*on the other hand*'), together with an arbitrary selection of two alternatives of the rich variety of pronunciation variants that are present in the corpus.

Table 5-6 Example of different pronunciations.

Orthography	*aan de andere kant*
Canonical transcription	`/an d@ And@r@ kAnt/`
Actual pronunciation 1	`/an d. Andr@ kAn./`
Actual pronunciation 2	`/An d. And@.. kAnt/`

Align uses a dynamic programming procedure to align two sequences of phonetic symbols. It computes two kinds of distance measures, one based on an articulatory feature representation of the transcription symbols, and one based on the number of substitutions, deletions and insertions observed between the two strings in question. During the alignment procedure, proper penalties for symbol substitutions are calculated in terms of articulatory features, such as place and manner of articulation, voice, lip rounding, length, etc. For deletions and insertions a fixed penalty is used. In addition to the feature based phonetic distance, *Align* also outputs a distance measure in the form of the percentage disagreement between the two sequences of symbols aligned. Percentage disagreement is the total number of differences between the two strings, divided by the number of segments in the canonical transcription:

$$\%disagreement = \frac{\#S + \#D + \#I}{\#phonemes} * 100\%$$

Although percentage disagreement might seem to be much coarser a measure than the feature based phonetic distance, we decided to use percentage disagreement in this study. The most important reason for doing so is that we expected that the bulk of the differences between canonical and observed pronunciations would consist of deletions in the observed pronunciations. All deletions obtain the same weight in the present version of *Align*. Moreover, results based on percentage disagreement would be easier to compare and replicate by other research teams.

5.3.3 Results

In the following sections we present the data concerning the actual pronunciation of the words contained in the N-grams. We first show how these pronunciations differ from their canonical representations. Next, we explain and motivate a further reduction of the set of N-grams under analysis for the more detailed comparison of pronunciation variants between words in what may be MWEs and the same words occurring in arbitrary contexts, and we present the quantitative results. Finally, the results of qualitative analyses of these pronunciations are presented.

N-gram pronunciation versus canonical

All the observed pronunciations of the 3-grams and 4-grams in Table 5-5 were aligned with the canonical representation of that specific N-gram. In the canonical representation no pronunciation variation due to context (cross-word processes) is modelled; only obligatory word internal phonological rules are applied. Although pronunciation variation due to cross-word context is very common in real speech, we choose to use this strict canonical transcription as reference material, because it is the only objective reference that can be used to generalize over contexts.

The discrepancy between the observed pronunciation and the canonical representation is expressed in percentage of substitutions, deletions and insertions relative to the number of phonemes in the canonical representations. In Table 5-7 the results of the alignment of all 3-grams and 4-grams are presented, separately for the six categories from Table 5-3 and expressed in an average percentage of disagreement (column 'total', subdivided into substitutions, deletions and insertions) together with the number of types belonging to each category. A detailed results table for each N-gram separately can be found in Appendix A3 and A4.

Table 5-7 Average percentage substitutions, deletions and insertions after alignment with canonical transcription.

	3-grams					4-grams				
	#type	%sub	%del	%ins	total	#type	%sub	%del	%ins	total
cat 1	32	13.89	9.14	0.47	23.50	9	15.37	13.79	1.54	30.70
cat 2	31	11.36	11.82	0.15	23.33	3	5.75	16.25	0.04	22.04
cat 3	4	11.46	15.21	0.70	27.36	2	20.33	8.66	0.00	28.99
cat 4	28	13.13	12.81	0.27	26.21	6	13.74	15.49	0.40	29.63
cat 5	1	6.00	10.00	0.00	16.00	1	3.57	15.00	0.00	18.57
cat 6	17	12.59	10.49	0.66	23.75	0	-	-	-	-

From Table 5-7 it can be seen that for all the 3-grams and 4-grams most of the differences between the canonical representation and the actual pronunciation are caused by deletions and substitutions of segments in the actual pronunciation. Only few insertions are observed. In quantitative terms this is precisely what one would expect: spontaneous speech is characterized by what could be considered as 'sloppy' pronunciation.

The dynamic programming algorithm used for alignment provides information not only on the number of discrepancies, but also on their nature. We found that the majority of phonemes that are deleted in the actual pronunciation of the N-grams are word final /t/, /n/ and /r/. Furthermore, many schwas, /@/, were deleted as well in both the 3-grams and the 4-grams. Most of the substitutions concern the reduction of full vowels in the canonical to schwas in the actual pronunciation. Many other substitutions involved the feature voice, where the unvoiced variant was most often found in the actual pronunciation. The few insertions observed seem to be related to processes that may be motivated by ease of articulation, such as homorganic glide insertion: insertion of /j/ or /w/ between two vowels (Booij, 1995), e.g. in the word 'zoiets' ('*something*'). The canonical transcription is /zoits/, but in the observed pronunciations the most frequent form is /zowits/. Thus, our data form a quantitative confirmation of the abundant presence of 'sloppy speech' phenomena that have been impressionistically described for spontaneously spoken Dutch (Ernestus et al., 2002).

From Table 5-7 it can also be seen that the total percentage disagreement is quite similar for all the categories. Therefore, it is not possible to pursue the analysis of differences between 'true' MWEs, stock phrases and coincidental frequent word sequences in depth in the remainder of this study.

Effect of contexts on pronunciation of words in N-grams

Although the number of N-grams with a sufficiently high frequency in the CGN *core corpus* (cf. Table 5-5) does not seem impressive, it is still far too high to allow a detailed comparison of pronunciation phenomena between words in N-gram context and in arbitrary contexts. The major cause of the problem is that it is not clear whether the percentage disagreement for individual words in an N-gram can be accumulated to provide a meaningful score for the complete sequence, without thorough analysis of the phenomena that caused the discrepancies in the first place. Therefore we decided to process data manually, which requires a further reduction of the data. Because we are interested in the potential effect of MWE status on pronunciation variation, we decided to select those N-grams from the corpus summarized in Table 5-5 which showed the highest degree of discrepancy between the actual pronunciation and the canonical reference. In this way we selected the ten 3-grams shown in Table 5-8 and the ten 4-grams in Table 5-9.

In addition to the N-grams shown in the tables, we also had to select occurrences of all words in these N-grams in 'comparable' arbitrary contexts. As explained in section 5.3.2 we defined 'comparable arbitrary context' in terms of the position in an arbitrary N-gram, with the only additional restriction that the neighbouring words must be different from the neighbours in the MWE N-gram. The number of other contexts for a word differs enormously between the words. For example, the word 'ware' ('*were*') occurs only once outside the context 'als het ware' ('*as it were*'), and the word 'een' ('*a*') from 'op een gegeven moment' ('*at a given moment*') occurs, of course, many more times. Each individual word has two collections of pronunciations, those found in the MWE context and those found in all other contexts. The same canonical transcriptions were used as a reference for the comparison of the actual pronunciations in the two context conditions.

Comparing the percentage disagreement observed for each word in the two context conditions gives the results displayed in Table 5-8 and Table 5-9. The percentage disagreement of an N-gram in one of the two contexts is the weighted total of the average percentages disagreement of the individual words in that specific N-gram. The individual percentage disagreement of a word is normalized for the frequency of occurrence, which is different in the two contexts and varies per word. The weighting for the summation of the individual percentages disagreement is determined by the number of phonemes of the word in the reference transcription. The expressions listed in column 1 are ranked according to the difference in percentage disagreement between the two conditions. A detailed results table for each word in the N-grams can be found in Appendix A5 and A6.

Table 5-8 Difference in percentages disagreement between two context conditions for words in 3-grams.

3-gram	category	%disagreement MWE context	%disagreement other context	difference
zoiets van ja	6	57.27	15.75	41.52
in ieder geval	2	37.17	12.26	24.91
af en toe	2	34.76	15.15	19.61
op die manier	2	31.94	12.99	18.95
't is natuurlijk	4	45.59	31.11	14.48
weet ik niet	1	29.22	21.52	7.7
dat is natuurlijk	4	34.62	28.76	5.86
hoe heet dat	1	30.43	24.95	5.48
ook helemaal niet	2	27.78	24.40	3.38
als 't ware	3	23.15	35.88	-12.73

Table 5-9 Difference in percentages disagreement between two context conditions for words in 4-grams.

4-gram	category	%disagreement MWE context	%disagreement other context	difference
dat vind ‘k ook	1	48.89	29.00	19.89
op een gegeven moment	2	47.13	27.91	19.22
dat maakt niet uit	1	42.42	26.49	15.93
dat is niet zo	1 and 4	40.00	28.47	11.53
of wat dan ook	3	31.54	22.10	9.44
‘k weet niet precies	4	28.57	22.73	5.84
dat weet ik niet	1	29.03	25.96	3.07
weet ik veel wat	3	26.45	25.08	1.37
dat weet ik nog	1	24.55	26.15	-1.60
als ‘t goed is	5	18.57	32.41	-13.84

The first observation that can be made from Table 5-8 and Table 5-9 is that selecting N-grams on the basis of their pronunciation yields mainly N-grams belonging to the categories that represent complete syntactic constituents. Although these categories were overrepresented (see Table 5-7) compared to the others, these results do confirm the intuition that there must be a relation between frequency of N-grams and syntactic constituency.

For both the selected 3-grams and 4-grams in Table 5-8 and Table 5-9 tests revealed that the differences in percentage disagreement between the two context conditions are significant (for 3-grams: $p = 0.010$ and for 4-grams: $p = 0.030$). Thus, it is safe to say that, on average, the pronunciation of words in the context of frequent N-grams differs more from the canonical form than the pronunciation of these words in arbitrary contexts. This finding also strongly suggests that many of the highly frequent N-grams in Table 5-8 and Table 5-9 qualify for the status of MWE, if not for another reason, then at least because of their effect on pronunciation.

Qualitative analyses

In order to get more insight into the type of pronunciation variation that characterizes these 20 frequent 3- and 4-grams, the differences between the transcriptions in the two context conditions were also analyzed on a qualitative level based on the output of *Align*. In Table

5-10 we show how many of these discrepancies were caused by deletions, substitutions and insertions.

Table 5-10 Average percentage disagreement (substitutions, deletions and insertions) for both context conditions.

average %	sub	del	ins	total
3-gram in MWE context	15.43	19.19	0.30	34.92
3-gram in other context	12.84	10.54	0.60	23.98
4-gram in MWE context	13.58	23.21	0.54	37.33
4-gram in other context	13.85	12.42	0.48	26.75

It is clear from this table that in both context conditions there are more deletions than insertions with respect to the canonical representations, which indicates that in both cases the actual pronunciations are reduced in comparison to their canonical reference. Since there are considerably more deletions in the condition 'MWE context', it is legitimate to conclude that in this case the pronunciation of the individual words is more reduced than in the condition 'other contexts'. However, to get a better understanding of the type of reduction that affects the individual words when they appear in the context of N-grams, it is important to look not only at the number of deletions, but also at possible relations between deletions in individual words. Specifically, we are interested in the possibility that in 'MWE context' the deletion of a cluster of phonemes occurs more often than in 'other contexts'. If deletion clusters are one of the specific phenomena for MWE contexts, they cannot be properly accounted for in the form of rewrite rules applied to individual words when generating a multi-pronunciation lexicon. To this end, we counted the number of deletion clusters of different length for all the words in the two context conditions (see Table 5-11).

Table 5-11 Distribution of deletion clusters of different sizes.

%	length 1	length 2	length 3	length 4
3-gram in MWE context	70.88	12.94	15.88	0.29
3-gram in other context	90.40	6.85	2.68	0.04
4-gram in MWE context	61.18	37.89	0.62	0.31
4-gram in other context	95.48	4.52	0.00	0.00

Table 5-11 clearly shows that the size and the distribution of deletion clusters are different in the two context conditions. In the condition 'MWE context' there are clearly more

deletion clusters of size 2, 3, and 4 than in the condition 'other contexts'. In other words, in the context of N-grams it is more common that sequences of two or three segments, therefore possibly whole syllables, are deleted. In addition, the fact that deletion clusters of a given size (i.e. 3 and 4 for 4-grams) are not found at all in the condition 'other contexts' seems to suggest that there are pronunciation variants that are unique for the 'MWE context' condition. Obviously, this is a point that deserves further investigation.

Qualitative analyses were also carried out for the data concerning the substitutions (cf. Table 5-12). In Table 5-10 we saw that the percentages of substitutions with respect to the canonical representation are similar in the two context conditions. Qualitative analyses of these substitutions also revealed that the processes underlying them are very similar. Table 5-12 shows that the most frequent substitutions concern processes such as voice assimilation and vowel reduction that are already known from the literature (Booij, 1995).

Table 5-12 Ten most frequent substitutions with percentage disagreement in both context conditions for 3-grams and 4-grams.

3-grams				4-grams			
MWE context		other context		MWE context		other context	
sub-type	%dis	sub-type	%dis	sub-type	%dis	sub-type	%dis
/t/-/d/	2.86	/t/-/d/	2.84	/t/-/d/	3.21	/t/-/d/	3.36
/k/-/g/	2.23	/d/-/t/	1.74	/k/-/g/	2.32	/k/-/g/	2.18
/v/-/f/	1.90	/k/-/g/	1.45	/v/-/f/	1.38	/d/-/t/	2.10
/E/-/@/	1.23	/s/-/z/	1.41	/A/-/@/	1.04	/A/-/@/	1.91
/I/-/@/	1.08	/A/-/@/	1.25	/d/-/t/	0.94	/I/-/@/	1.01
/d/-/t/	0.93	/v/-/f/	1.03	/E/-/@/	0.94	/s/-/z/	0.77
/a/-/@/	0.89	/I/-/@/	0.77	/p/-/b/	0.69	/z/-/s/	0.48
/a/-/A/	0.63	/a/-/A/	0.36	/s/-/z/	0.49	/v/-/f/	0.31
/f/-/v/	0.52	/a/-/@/	0.29	/n/-/N/	0.49	/n/-/m/	0.30
/z/-/s/	0.41	/E/-/@/	0.25	/e/-/@/	0.35	/A/-/a/	0.27
total	12.68		11.40		11.85		12.69

5.4 Discussion

The analysis of frequent N-grams showed that a very large proportion (21%) of the words in the spontaneous speech in the CGN corpus is part of word sequences that occur frequently. This highly repetitive and predictable nature of extemporaneous speech

deserves more attention in the future than it has received in the past. Furthermore, while compiling the set of frequent N-grams, we also found that there are quite a number of N-grams which occur frequently in very specific communicative settings and not at all in other settings. Whether this finding is coincidental or systematic can only be determined by comparing and analyzing more and larger spoken corpora than just the CGN.

In the CGN we have observed a tendency for frequent N-grams to consist of complete syntactic clauses, or at least opening part of a clause. Although this finding is intuitively plausible, we still need further research to understand its implications for psycholinguistics and speech technology. The results presented in section 5.3.3 clearly indicate that for all the words in the N-grams investigated the actual pronunciation is reduced with respect to its canonical representation. The amount of reduction in pronunciation is mainly caused by the fact that many segments in the canonical representation appear to be deleted in the actual pronunciation. In addition, analyses of the substitutions observed reveal that many of these also concern reduction processes: i.e. substitutions of full vowels in the canonical transcriptions by schwas in the actual pronunciations. So, these results confirm those of previous investigations which have shown that in spontaneous casual speech words may be highly reduced (Ernestus et al., 2002; Keating, 1998; Kohler, 1990).

However, in our study we wanted to determine whether this amount of reduction is characteristic of spontaneous speech across the board, or whether it is related to specific contexts, in particular those of frequent N-grams. To answer this question we examined the pronunciation variants of the same words in the context of N-grams and in all remaining contexts. The results of these analyses, presented in section 5.3.3, make it clear that for almost all the words investigated it holds that the degree of reduction is higher when these words appear in the context of frequent N-grams as opposed to when they appear in any other context. Moreover, analyses of the distribution of deletions reveal that in the context of frequent N-grams deletions tend to be more grouped together than in the other contexts, indicating that sometimes whole syllables are deleted in N-grams. Finally, the fact that the clustering pattern of deletions is different in the two context conditions and that certain cluster types are not found outside frequent N-grams indicates that 'MWE-like' N-grams probably contain unique pronunciation variants. These findings suggest that, at least for the purpose of pronunciation modelling, it is necessary to add a number of frequent N-grams with their characteristic pronunciation variants to the (pronunciation) lexicon. This may be a better solution than indiscriminate addition of all the pronunciation variants observed to the individual words in the lexicon, which, as shown in Kessens et al (1999), is counter-productive.

The most important reason to start the research reported in this paper was to determine whether these MWEs and their pronunciation variants require special treatment in

automatic speech recognition (ASR) and automatic phonetic transcription (APT). Previous research has shown that modelling pronunciation variation can be beneficial for both APT and ASR: for APT because the quality of the resulting transcriptions can be improved (Binnenpoorte et al., 2004; Schiel, 1999); and for ASR, because the word error rates can be reduced (Strik & Cucchiarini, 1999). In ASR research it has also been shown that if too many variants are added, word error rates increase again. Specific modelling of pronunciation variation in MWEs has been studied in the field of ASR, but, as far as we know, not in the field of APT. In ASR, MWEs are referred to as phrases, word tuples, multiword units, or multiwords. Different criteria are used to select, usually a small number of, MWEs. Adding these MWEs and their pronunciation variants to the lexicon usually reduces word error rate. In general, the main goal of these studies is to reduce word error rate, and, consequently, no detailed study of pronunciation variation of MWEs is carried out. In our study we did examine the type of pronunciation variation that characterizes a selected number of frequent MWEs and found that these exhibit uncommon pronunciation patterns that are not found in other contexts. We therefore suggest that these MWEs be included as lexical entries in the pronunciation lexicons employed in ASR and APT, because in both cases this is likely to improve the performance of the system.

5.5 Conclusions and perspectives for future research

In this paper we have presented an exploratory study of MWEs in spontaneous speech in which focusses on the pronunciation of MWEs in relation to ASR and APT. We have shown that the words composing the MWEs investigated do indeed exhibit different pronunciation patterns in the MWE context than in other contexts. This provides evidence for the fact that these MWEs require special treatment in ASR and APT.

The results of our study suggest that phonetically transcribed corpora are a valuable source for research into phenomena and problems that affect the performance of ASR and APT for conversational speech and that have so far been elusive. However, the practical problems encountered in this study also make it clear that eventually we will need phonetically transcribed corpora of unprecedented size. Therefore, it is essential to continue the research aimed at developing accurate automatic phonetic transcriptions of speech recordings. The results obtained with our medium size corpus already show a number of promising directions for that research.

Future research could also profit from the application of shallow syntactic parsing to the classification of N-grams that we have performed on the basis of the orthography alone. More detailed information about the type and the degree of completeness of the syntactic

constituent formed by frequent N-grams should help in selecting the word sequences that are candidates for inclusion in a MWE lexicon.

Adding information about prosody, if only in the form of the strength of the juncture between adjacent words, is an obvious extension of the work reported in this paper. It seems evident that the presence of clear phonetic boundaries between adjacent words prevents the deletion of large phoneme clusters across the boundary. However, here too one will need large corpora with accurate transcriptions to support the research.

Acknowledgements

The authors would like to thank Nelleke Oostdijk, Peter-Arno Coppen and Bill Fletcher and the three anonymous reviewers for their careful and constructive comments on earlier versions of this paper.

GENERAL DISCUSSION AND CONCLUSIONS

CHAPTER 6

6.1 Discussion and conclusions

6.1.1 Transcription procedure for large speech corpora

Many large general speech corpora have been compiled in the last decade (Cresti et al., 2004; Furui et al., 2000; Godfrey et al., 1992; Hess et al., 1995; Pitt et al., 2005; Wissing et al., 2004), and probably more will be in the near future, to serve the needs for data in the fields of linguistics, phonetics, and language and speech technology. For all speech corpora an orthographic transcription of the speech signal is indispensable. For most applications it will remain necessary to produce this orthographic transcription by human labour. Some types of speech research require additional, more detailed annotation of the speech signal, such as a broad phonetic transcription. A complete human-made phonetic transcription of a large speech corpus is practically impossible, since these transcriptions are very time-consuming and therefore prohibitively expensive. Recourse to (semi-)automatic phonetic transcription procedures is inevitable.

The challenge for automatic phonetic transcription lies in capturing the variability in human speech. The pronunciation of a word can be different each time the word is uttered, either by another speaker or by the same speaker. This variability is especially large and difficult to predict in more conversational, spontaneous settings, where speakers tend to speak less carefully (Engstrand, 1992; Kohler, 1998; Swerts et al., 2003). While there are other ways to produce automatic phonetic transcriptions, the transcription procedure used in the CGN project, as described in this thesis, followed a method in which the best matching variant is chosen from a closed list of possible pronunciation variants. In order to be able to automatically transcribe human speech with this procedure, one has to know how to generate that list of variants, preferably with their relative frequencies of occurrence. Most of the rule-based information that is available for the pronunciation of Dutch concerns processes occurring in laboratory, read speech (Booij, 1995; Cutler, 1998). With respect to more casual speech, the knowledge is limited to pronunciation phenomena observed in business negotiations (Ernestus, 2000) and existing knowledge only describes in general terms which phonological processes can occur in certain phonemic contexts. However, quantitative information that specifies which processes can be applied exactly where and when is not yet available, for any style of spoken Dutch. Moreover, the CGN contains examples of several speech styles, such as television commentary, conferences and lectures that have not yet been investigated in any detail.

In situations where phonological knowledge is lacking and large amounts of phonetically transcribed data from which statistical knowledge could be extracted are not

yet available, the bootstrap procedure described in chapter 2 offers a solution. This bootstrap procedure consists of four cycles of increasing complexity with respect to transcription generation effort. Each cycle contains four stages, i.e. generation, validation, diagnosis and remedy. Through continuous quality control and error analysis on two small independent samples, quantitative information on several frequent phonological processes in various speech styles could be extracted. Subsequent deployment of the newly obtained information in the generation of a new automatic phonetic transcription showed that the information was relevant and reliable in the sense that it gave rise to improved transcription quality. Although the procedure seems somewhat trivial for well-described languages, for the exploration of new speech corpora containing languages and regional variants for which phonological knowledge is lacking or is at least not documented, the procedure is all the more useful.

To conclude, an iterative automatic phonetic transcription procedure, in which more knowledge about pronunciation phenomena is acquired in each iteration, is especially useful in circumstances in which a new large speech corpus needs to be transcribed that contains speech styles for which qualitative and quantitative phonetic knowledge is not yet available.

6.1.2 Data-driven knowledge extraction from existing speech corpora

In the automatic transcription procedure in chapter 2 we confined ourselves to rule-based pronunciation variant generation. Where the variation in pronunciation in read speech is rather predictable and can be covered with rewrite rules, at least for Dutch, in spontaneous speech the variation is much richer. Words can be heavily reduced; even complete syllables can be deleted (Fosler-Lussier & Morgan, 1999). Modelling this capricious variation is virtually impossible with a manageable set of rewrite rules, which emphasises the need to explore the possibilities of other techniques for the generation of a list of plausible pronunciation variants for spontaneous speech.

In 2004 the CGN was released, which is the first available corpus that contains large amounts of hand-transcribed spontaneous Dutch. The availability of this corpus opens up opportunities to study pronunciation phenomena in spontaneous speech in a data-driven way (cf. e.g. Cremelie & Martens, 1999; Kessens et al., 2003; Schiel et al., 1998). In chapter 3 we applied an enumeration technique (Strik & Cucchiarini, 1999) and listed all possible pronunciation variants in a lexicon together with their prior probabilities as observed in the data. In order to test the value of the enriched lexicon we had a small sample automatically transcribed. By means of forced recognition the best matching variant was chosen from the lexicon. The prior probability of each variant was stored in a

class-based language model in order to put constraints on the forced recognition system. The resulting transcription of the small sample appeared to be improved compared to the transcription of spontaneous speech for which the variants were generated by means of rewrite rules only (chapter 2).

The list of data-extracted pronunciation variants and their frequency information is potentially also a very useful knowledge source for ASR applications. However, for both automatic transcription and ASR the enumeration method cannot suffice with just listing variants observed in some corpus. It is clear that the list containing the pronunciation variants is dependent on the occurrence of words and their pronunciations in the corpus from which they were extracted. For unseen words and additional plausible pronunciation variants other than those actually occurring in the corpus, an additional rule-based variant generation procedure needs to be applied.

Given the potentials for automatic phonetic transcription as reported in chapter 3, we can conclude that large spontaneous speech corpora containing good quality manual phonetic transcriptions are particularly useful for the extraction of new knowledge, i.e. a list of pronunciation variants and their frequency of occurrence, on phenomena in spontaneous speech. This new knowledge can subsequently be deployed for an improved automatic phonetic transcription of future spontaneous speech corpora.

The initial exploration of the hand-transcribed corpus of spontaneous Dutch described in chapter 3 suggested that it might be fruitful to search for systematic phenomena beyond the phoneme level. We aimed at gathering additional information, both qualitative and quantitative, about the occurrence of extremely reduced pronunciations in spontaneously spoken Dutch. Bell et al. (2003) examined American English spontaneous speech from the Switchboard corpus. They found among other things that words that are contextually predictable are produced with less articulatory effort. In the research described in chapter 5 we found similar patterns for spontaneous Dutch. Words occurring in frequently found contexts, which were referred to as Multiword Expressions, may show extreme reductions. Occurrences of the same words outside the multiword context showed significantly fewer reduction patterns compared to a canonical representation. This finding provides support for the fact that these expressions probably need special treatment in future attempts at automatic phonetic transcription and in ASR applications that must deal with spontaneous speech.

To conclude, the currently available hand-transcribed corpus has proven to be very useful for an exploration of spontaneous speech. Future corpus compilation projects aiming at phonetic transcriptions of spontaneous data can already make use of the results obtained so far.

6.1.3 Manual transcription procedure

In the experiments reported in this thesis the quality of both machine and hand-made transcriptions was measured by means of a symbol-to-symbol comparison with a reference transcription. We have argued that the reference transcription should be made in consensus mode and from scratch by at least two experts (cf. Shriberg et al., 1984). Owing to constant negotiation on each symbol, transcriber-specific errors are less likely to be present in the resulting consensus transcription. This is what makes a consensus transcription more suitable as reference than e.g. a majority vote transcription derived from a number of individually made transcriptions, at least when these are made by editing an example transcription.

In chapter 4 we aimed at evaluating the manual transcription procedure as it was applied in the CGN and many other corpora. In order to do this we examined the hand-made transcriptions made by editing a given example transcription. An experiment was set up in which CGN transcribers were asked to transcribe a small sample of the CGN corpus according to the CGN protocol. In establishing transcription quality we measured intra-transcriber agreement, as is done in Eisen (1993), Greenberg et al. (1996), Kikuchi and Maekawa (2003) Raymond et al. (2002), and Wesenick and Kipp (1996), as well as the distance between the individual transcriptions and the reference transcription.

Owing to the consensus procedure, the reference transcription can be considered as a good approximation of ground truth, to which individually made transcriptions can be compared in terms of a distance measured as the proportion of symbols that are different between the two transcriptions. Since in the reference transcription transcriber-typical extremities are likely to be absent, we can consider this reference transcription as some sort of middle estimate, around which the various individually made transcriptions are located. Contrary to what one might expect, the distance between the reference transcription and each of the individual transcriptions was larger than the distance between the various individual transcriptions (inter-transcriber agreement). This observation suggests that the individually made transcriptions contained fewer subjective elements and idiosyncrasies as could be expected when these transcriptions were made from scratch without any example transcription. It is highly likely that the inter-transcriber agreement scores are artificially high due to the given example transcription that needed to be verified by the individual transcribers. This makes inter-transcriber agreement as a sole measure of transcription quality less suitable if different transcribers are required to check and correct the same example transcriptions. The comparison with a consensus transcription showed that caution is required regarding interpreting inter-transcriber agreement scores when transcriptions are made with a verification and correction procedure.

6.1.4 Transcriptions of read speech

In the context of large speech corpora automatic phonetic transcriptions constitute an interesting alternative for human-made phonetic transcriptions. Automatic transcriptions can facilitate the task of the human transcriber by serving as a high quality example transcription. It may even be the case that automatic transcriptions can replace human transcriptions. Clearly, the replacement of human transcribers is only feasible if the automatic transcriptions are good enough. Good enough here means 'comparable to the quality obtained with hand-made transcriptions of large speech corpora by students working under time pressure'. In chapter 2 we have argued for a threshold, expressed in maximum proportion of phoneme disagreement, below which the automatic transcription is considered to be good enough. We took into account the advantages of a fast and cheap automatic procedure opposed to an expensive manual procedure.

Relating the quality of automatic transcriptions as obtained in chapter 2 to human-made transcriptions for read speech showed that good quality automatic phonetic transcriptions on a broad phonetic level can be obtained by concatenating canonical forms on which two cross-word processes were applied, i.e. assimilation of voice and degemination.

In chapter 4 the quality of the transcriptions obtained by manually correcting automatically generated example transcriptions was assessed. For all speech styles involved in the experiment, the human-corrected transcription outperformed the automatic transcription. However, the degree of improvement was not equal for the different speech styles. We found that for read speech this improvement was actually rather small; the automatic example transcription differed in 10.5% of the phonemes from the reference transcription, whereas the manual transcriptions still differ in 6.3% of the phonemes from the same reference. Adding this to the observation that human transcribers corrected on average about 10% of the symbols in the automatic transcription that was set as the example transcription to check and correct, it can be concluded that quite some effort can be saved by omitting human corrections of an example transcription of read speech.

Based on the above, the following conclusion can be drawn: good quality broad phonetic transcription for read speech can be obtained fully automatically by using relatively simple techniques. By omitting human correction of an automatically generated example transcriptions of carefully produced speech, a lot of time and thus money can be saved that can be allocated for the benefit of phonetic transcriptions of speech styles for which larger deviations from a canonical representations are to be expected.

6.1.5 Transcriptions of spontaneous speech

Whereas for read speech already good quality automatic broad phonetic transcriptions can be obtained with limited resources, for more spontaneous speech styles this is certainly not the case. In chapters 2 and 3 we applied a forced recognition technique, where an ASR had to choose the best matching pronunciation variant from a closed list of candidates. The way in which the pronunciation variants were generated, by means of rules or by enumeration, differed between the two experiments. In chapter 2 we found a modest improvement in transcription quality with rule-based techniques. In chapter 3 the improvement obtained with the enumeration technique was somewhat larger. But still, both techniques applied in the two chapters did not give satisfactory results in the sense that the automatic transcriptions of spontaneous speech were good enough to make the human correction phase superfluous.

Although it is clear that human transcribers deviate more from the reference transcription as the speech style becomes less prepared and controlled, for spontaneous speech manually corrected example transcriptions yield much better results than our automatic procedure. Spontaneous speech seems to be intrinsically more difficult to transcribe than read speech because of various factors, such as the technical quality of the speech signal itself, the degree of deviation from a canonical pronunciation, and the lower level of intelligibility, overlapping speech fragments, and disfluencies (see also González et al., 2004). The disappointing result of our automatic procedure is partly due to the characteristics of spontaneous speech and partly due to the lack of relevant phonological knowledge. Furthermore, the ASR we used in chapters 2 and 3 is suboptimal and cannot, for instance, easily detect shortened phonemes, while shortening is a frequently applied process in spontaneous speech (Swerts et al., 2003). Adapting the ASR and, more importantly, increasing the phonological knowledge on spontaneous speech, might result in improved automatic phonetic transcription. Manual transcriptions of large amounts of spontaneous speech already proved to be successful for acquiring additional knowledge, so more hand-made transcriptions are needed in the future.

To conclude, for spontaneous speech human transcriptions are still the best option, although improved automatic techniques, together with a better understanding of the phonological processes underlying spontaneous speech, are likely to approximate human transcription quality for spontaneous speech styles in the future.

6.2 Future work

Clearly, there is still a lot of work that can be done to improve automatic phonetic transcriptions of spontaneous spoken Dutch. Enhanced automatic transcriptions should ultimately be able to make human-made transcriptions superfluous. Fully automatic phonetic transcriptions have a huge cost and time advantage over human-made transcriptions. However, before this can be achieved, thorough analyses of pronunciation processes in spontaneous speech are necessary. The basic requirements for such analyses are large amounts of well-transcribed spontaneous speech recordings. We have shown that a medium-sized hand-transcribed corpus of spontaneous Dutch already served as a useful resource to obtain knowledge on spontaneous speech processes. But still, larger data sets, perhaps recorded in different situational settings and with different groups of speakers, are necessary to further complete the knowledge on pronunciation processes and to improve automatic transcription procedures. These new data sets, in turn, require good quality phonetic transcriptions, which ultimately should be made automatically, thus ensuring that new speech corpora become available more rapidly and at lower costs.

The implementation of variant-based data-driven pronunciation variation modelling, as described in chapter 3, for the benefit of improving automatic transcriptions has some limitations. The most obvious restriction lies in the fact that this approach can only be adopted under the condition that a large manually generated phonetically annotated corpus is available. Another disadvantage is that only pronunciation variants that actually occur in the source data can be included in a pronunciation lexicon. Words that do not occur in the source data, but that do occur in the data to be transcribed, could only be assigned the canonical transcription in the lexicon. It is obvious that in this way pronunciation variants that may be more plausible than the canonical variant are not included in the lexicon. Moreover, the list of variants is critically dependent on the data, and more specifically on the context in which a word occurred. Cross-word phonological processes, such as voice assimilation and degemination, are frequent phenomena in Dutch running speech. Obviously in a restricted training set (source data) not all contexts in which a word can appear can be accounted for, so cross-word processes for unseen contexts are not modelled either. Combining the variant-based method and a rule-based method (either knowledge-based or data-driven), could help to overcome the limitations imposed by unseen variants and unseen contexts. This could be a promising path towards better automatic transcriptions of spontaneous speech (cf., Riley et al., 1999).

Although our attempt to gather new knowledge on spontaneous speech phenomena as described in chapter 5 was successful, it is only just a beginning that requires more detailed research and the actual application of the newly obtained knowledge for the benefit of automatic transcriptions of spontaneous speech. Other factors that influence or predict certain pronunciation phenomena should be included in future explorations, such as prosody, syntactic and semantic information. Moreover, larger data sets than the one used in chapter 5 could potentially reveal additional pronunciation patterns in spontaneous Dutch. However, in this respect it is not only just the data, it is also the method by which new knowledge can be extracted. In order to process large resources, automatic techniques for the efficient analysis of the data need to be developed. With regard to detecting relevant deviant pronunciation patterns in spontaneous speech, Strik et al. (2005) suggest some measures, such as frequency of occurrence, relative and absolute string length difference between the canonical representation and the hand-corrected transcription, and speech rate. Future research aimed at developing additional measures to detect patterns of deviant pronunciations is needed.

Hand-transcribed speech corpora are essential knowledge sources for the study of pronunciation phenomena. Since many conclusions from the present study are based on the human-made transcriptions, the quality of these transcriptions was measured in chapter 4. We have come to the conclusion that in evaluating the verification and correction procedure that is often followed in large speech corpus projects caution is required in interpreting inter-transcriber agreement scores as quality indicators. It was observed that the example transcription is likely to bias the human-made transcriptions. However, in order to isolate and quantify the effect of the given example transcription on agreement scores, a new experiment must be set up. In such an experiment, the transcribers would have to transcribe a speech sample, first from scratch and than re-transcribe it with an example transcription. The possible bias effect can then be measured if the inter-transcriber agreement scores of the two transcription experiments are compared. Nevertheless, before interpreting the agreement scores, one should keep in mind that leaning effects of re-transcribing the same material can complicate the design of such an experiment. Consequently, quantifying possible bias effects of a given example transcription is a difficult task.

6.3 Final remarks

High quality speech corpora are extremely useful for language and speech technologists and researchers. Large speech corpora containing speech material recorded in real-life

situations, such as telephone conversations, interviews, etc., offer the possibility to study all kinds of phenomena in real-life speech, for instance, turn-taking management in a dialogue, the syntax of spoken language, prosody in telephone conversations, and so on. For the central research topic in this thesis – phonetic transcriptions of large speech corpora – the large speech corpora that are now available appear to be extremely valuable in various respects. These large speech corpora allow us to study pronunciation phenomena for the benefit of improving pronunciation modelling, and for the aim of developing more refined and more adequate methods for transcription generation and transcription assessment. In this way, phonetic transcription research does not only profit from large collections of speech, but also contributes to the future development of new, large and more comprehensive speech corpora for which phonetic transcriptions need to be generated.

BIBLIOGRAPHY

Baayen, R.H., Piepenbrock, R., Gulikers, L. (1995). *The CELEX Lexical Database – release 2*. [CD-ROM] Available at Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Ball, M.J., Rahilly, J. (2002). Transcribing disordered speech: The segmental and prosodic layers. *Clinical Linguistics & Phonetics, Vol. 16 (5)*, pp. 329-344.

Bell, A., Jurafsky, D., Fosler-Lussier E., Girand, C., Gregory, M., Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America (JASA) 133 (2)*, pp 1001-1024.

Berg, R. van den (1988). *The perception of voicing in Dutch two-obstruent sequences*. PhD thesis, Katholieke Universiteit Nijmegen.

Bernstein, J., Taussig, K., Godfrey, J. (1994). Macrophone: An American English telephone speech corpus for the Polyphone project. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, Adelaide, Australia, pp. 81-83.

Beulen, K., Ortmanns, S., Eiden, A., Martin, S., Welling, L., Overmann, J., Ney, H. (1998) Pronunciation modelling in the RWTH large vocabulary speech recognizer. *Proceedings of the ESCA Workshop Modeling Pronunciation Variation for Automatic Speech Recognition,* pp. 13-16.

Biber, D., Johannson, S., Leech, G., Conrad, S., Finegan, E. (1999) *The Longman Grammar of Spoken and Written English*, Longman, Harlow, Essex. pp. 987-1036.

Binnenpoorte, D. (2002). Protocol voor manuele verificatie van automatisch gegenereerde woordsegementaties. (in Dutch) [http://www.tst.inl.nl/cgndocs/wrd_prot_nl.pdf].

Binnenpoorte, D., Goddijn, S., Cucchiarini, C. (2003) How to Improve Human and Machine Transcriptions of Spontaneous Speech. *Proceedings of ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, Japan, pp. 147-150.

Binnenpoorte, D., Cucchiarini, C. (2003). Phonetic transcription of large speech corpora: How to boost efficiency without affecting quality. *Proceedings of 15th International Congress of Phonetic Sciences (ICPhS'03)*, Barcelona, Spain, pp. 2981-2984.

Binnenpoorte, D., Cucchiarini, C., Strik, H., Boves, L. (2004) Improving Automatic Phonetic Transcription of Spontaneous Speech through Variant-Based Pronunciation Variation Modelling. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, Lisbon, Portugal, pp. 681-684.

Binnenpoorte, D., Cucchiarini, C., Boves, L., Strik, H. (2005). Multiword expressions in spoken language: an exploratory study on pronunciation variation. *Computer Speech and Language,19(4)*, pp. 433-449.

Binnenpoorte, D., Cucchiarini, C., Boves, L. (submitted). Measuring phonetic transcription quality in large speech corpora. Submitted to *Language Resources and Evaluation*.

Booij, G. (1995). *The phonology of Dutch.* Clarendon Press, Oxford.

Boves, L., Oostdijk, N.H.J. (2003). Spontaneous speech in the Spoken Dutch Corpus. *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*. 14-16 April, Tokyo, Japan, CDROM.

Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R. (1992) Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18 (4), pp. 467-480.

Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H., Martens, J.-P., Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, Spain, pp. 779-785.

Chang, S., Shastri, L., Greenberg, S. (2000). Automatic phonetic transcription of spontaneous speech (American English). *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*, Beijing, China, pp. 330-333.

Coussé, E., Gillis, S., Kloots, H., Swerts, M. (2004). The influence of the labeller's regional background on phonetic transcriptions: Implications for the evaluation of spoken language resources. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, Lisbon, Portugal, pp. 1447-1450.

Cremelie, N., Martens, J.P. (1997). Automatic rule-based generation of word pronunciation networks. *Proceedings of Eurospeech '97*, Rhodes, Greece, pp. 2459-2462.

Cremelie, N., Martens, J.-P. (1999). In search of better pronunciation models for speech recognition. *Speech Communication 29*, pp. 115-136.

Cresti, E., Bacelar, F., Sandoval, A.M., Veronis, J., Martin, P., Choukri, K. (2004). The C-ORAL-ROM CORPUS. A multilingual resource of spontaneous speech for Romance languages. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, pp. 575-578.

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study*. PhD thesis, Katholieke Universiteit Nijmegen.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clinical Linguistics & Phonetics, Vol. 10 (2)*, pp. 131-155.

Cucchiarini, C., Strik H. (2003) Automatic phonetic transcription: An overview. *Proceedings of 15th International Congress of Phonetic Science (ICPhS'03)*, Barcelona, Spain, pp. 347-350.

Cutler, A. (1998). The recognition of spoken words with variable representations. *Proceedings of ESCA Workshop on Sound Patterns of Spontaneous Speech 1998*, Aix-en-Provence, France, pp. 83-92.

Daelemans, W., Bosch, A. van den (2001). TreeTalk: Memory-based word phonemisation. In Damper, R. I. (Eds.): *Data-Driven Techniques in Speech Synthesis.* Kluwer Academic Publishers, pp. 149-172.

Demuynck, K., Laureys, T., Gillis, S. (2002). Automatic generation of phonetic transcriptions for large speech corpora. *Proceedings of International Conference on Spoken Language Processing (ICSLP'02)*, Denver, USA, pp. 333-336.

Demuynck, K., Laureys, T., Wambacq, P., Van Compernolle, D. (2004). Automatic phonemic labeling and segmentation of spoken Dutch. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, pp. 61-64.

Eisen, B. (1993). Reliability of speech segmentation and labelling at different levels of transcription. *Proceedings of Eurospeech '93*, Berlin, Germany, pp. 673-676.

Engstrand, O. (1992). Systematicity of phonetic variation in natural discourse. *Speech Communication 11, issues 4 – 5*, pp. 337-346.

Ernestus, M. (2000). *Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface*. PhD thesis, LOT, Utrecht.

Ernestus, M., Baayen, H., Schreuder, R. (2002) The Recognition of Reduced Word Forms. *Brain and Language*, 81. pp. 162-173.

Finke, M., Waibel, A. (1997) Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. *Proceedings of EuroSpeech-97*, Rhodes, Greece. pp. 2379-2382.

Fosler-Lussier, E., Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication 29*, pp. 137-158.

Furui, S., Maekawa, K., Isahara, H. (2000). A Japanese national project on spontaneous speech corpus and processing technology. *Proceedings of ISCA ITRW ASR2000*, Paris, France, pp. 244-248.

Gillis, S. (2001). Protocol voor brede fonetische transcriptie. (in Dutch) [http://www.tst.inl.nl/cgndocs/fon_prot.pdf].

Goddijn, S. (2003). Brede fonetische transcripties in het Corpus Gesproken Nederlands. (in Dutch) In: *LINK 14(1)*, pp. 9-13.

Goddijn, S., Binnenpoorte, D. (2003) Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. *Proceedings of 15th International Congress of Phonetic Science (ICPhS'03)*, Barcelona, Spain, pp. 2981-2984.

Godfrey, J.J., Holliman, E.C., McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'92)*, San Francisco, USA, pp. 517-520.

Goedertier, W., Goddijn, S. (2000). Protocol voor orthografische transcriptie. (in Dutch) [http://www.tst.inl.nl/cgndocs/ort_prot.pdf].

Goedertier, W., Goddijn, S., Martens, J.-P. (2000). Orthographic transcription of the Spoken Dutch Corpus. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece, pp. 909-914.

González Ledesma, A., De la Madrid Heitzmann, G., Alcántara Plá, M., De la Torre Cuesta, R., Moreno Sandoval, A. (2004). Orality and difficulties in the transcription of spoken corpora. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, CDROM.

Greenberg, S., Hollenback J., Ellis D. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard Corpus. *Proceedings of the International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, pp. S32-35.

Greenberg, S. (1998). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Proceedings of the ESCA workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, Rolduc, The Netherlands, pp. 47-56.

Gussenhoven, C. (1981). Voiced fricatives in Dutch: sources and present-day usage. *Proceedings Institute of Phonetics Nijmegen 5*, pp. 84-95.

Gussenhoven, C. (1992). Illustrations of the IPA: Dutch. *Journal of International Phonetic Association, 22,* pp. 45-47.

Hawkins, S. (2002) Roles and representations of systematic fine phonetic detail in speech understanding, *Journal of Phonetics*, *Vol. 31*, pp. 373-405.

Hess, W., Kohler, K.J., Tillman, H.-G. (1995). The Phondat-Verbmobil speech corpus. *Proceedings of Eurospeech'95*, Madrid, Spain, pp. 863-866.

Heuvel, H. van den, Boves, L., Moreno, A., Omologo, M., Richard, G., Sanders, E. (2001). Annotation in the SpeechDat projects. *International Journal of Speech Technology, 4*, pp. 127-143.

Heuvel, H. van den, Cucchiarini, C. (2001). /r/-deletion in Dutch: rumours or reality? In: H. van de Velde & R. van Hout (Eds.): *r-atics: sociolinguistic, phonetic and phonological characteristics of /r/*. Special issue of *Etudes & Travaux*, ILVP, Bruxelles, pp. 185-198.

Hoste, V., Daelemans, W., Gillis, S. (2004). Using rule induction techniques to model pronunciation variation in Dutch. *Computer Speech and Language, 18(1)*, pp. 1-23.

Keating, P. (1998) Word-level phonetic variation in large speech corpora. A. Alexiadou, N. Fuhrhop, U. Kleinhenz, & P. Law (Eds.), *ZAS papers in linguistics, vol. 11*. Berlin: Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung. pp. 35-50.

Kessens, J.M., Wester, M., Strik, H. (1999). Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. *Speech Communication 29*, pp. 193-207.

Kessens, J.M. (2002). *Making a difference: on automatic transcription and modelling of Dutch pronunciation variation for automatic speech recognition*. PhD thesis, Katholieke Universiteit Nijmegen.

Kessens, J.M., Cucchiarini, C., Strik, H. (2003). A data-driven method for modeling pronunciation variation. *Speech Communication 40*, pp. 517-534.

Kessens, J.M., Strik, H. (2004). On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions. *Computer Speech and Language 18*, pp. 123-141.

Kikuchi, H., Maekawa, K. (2003). Performance of segmental and prosodic labeling of spontaneous speech. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, CD-ROM.

Kipp, A., Wesenick, M-B., Schiel, F. (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. *Proceedings of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, pp. 106-109.

Kipp, A., Wesenick, M-B., Schiel, F. (1997). Pronunciation modelling applied to automatic segmentation of spontaneous speech. *Proceedings of Eurospeech '97*, Rhodes, Greece, pp. 1023-1026.

Kissine, M., Van de Velde, H., Hout, R. van (2003). An acoustic study of standard Dutch /v/, /f/, /z/ and /s/. In: Fikkert, P. and Cornips, L. (Eds.): *Linguistics in the Netherlands 2003*. Amsterdam: John Benjamins, pp. 93-104.

Knowles, G. (1994). Annotating large speech corpora: building on the experience of Marsec. *Journal of Linguistics*, *13*, pp. 87-99.

Kohler, K.J. (1990) Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, Dordrecht, Kluwer. pp. 69-92.

Kohler, K.J. (1998). The phonetic manifestation of words in spontaneous speech. *Proceedings of ISCA Workshop on Sound Patterns of Spontaneous Speech (SPoSS) 1998*, La Baume-les-Aix, France, pp. 13-22.

Koster, C.H.A. (2004) Transducing Text to Multiword Units. *Proceedings MEMURA 2004 workshop*, Lisbon, Portugal. pp. 31-38.

Lamel, L., Kassel, R.H., Seneff, S. (1986). Speech database development: design and analysis of the acoustic-phonetic corpus. *Proceedings DARPA Speech Recognition Workshop*, pp. 100-109.

Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press, Cambridge.

Marsi, E. (2003). Prosodische annotatie van het CGN. In: *LINK 14(1),* pp. 18-22.

Martens, J.P., Binnenpoorte, D. Demuynck, K., van Parys, R., Laureys, T., Goedertier, W., Duchateau, J. (2002). Word Segmentation in the Spoken Dutch Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, Spain, pp 1432-1437.

Nivre, J., Nilsson, J. (2004) Multiword Units in Syntactic Parsing. *Proceedings MEMURA 2004 workshop*, Lisbon, Portugal. pp. 39-46.

Nunberg, G., Sag, I.A., Wasow, T. (1994) Idioms, *Language*, *70*. pp. 491-538.

Odijk, J. (2004) Reusable Lexical Representations for Idioms. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. pp. 903-906.

Oostdijk, N.H.J. (2000). The Spoken Dutch Corpus. Outline and first evaluation. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece. Vol. 2: pp. 887-894.

Oostdijk, N.H.J. (2002). The design of the Spoken Dutch Corpus. In: P. Peters, P. Collins and A. Smith (Eds.): *New Frontiers of Corpus Research*. Amsterdam: Rodopi, pp. 105-112.

Oostdijk, N.H.J. (2004a). Het Corpus Gesproken Nederlands. [http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/formats/text/fon.htm]

Oostdijk, N.H.J. (2004b). Het Corpus Gesproken Nederlands. [http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/lexicon/README_cgnlex7.htm]

Os, E. den, Boogaart, T.I., Boves, L., Klabbers, E. (1995). The Dutch Polyphone corpus. *Proceedings of Eurospeech'95*, Madrid, Spain, pp. 825-828.

Pallett, D.S. (2003) A look at NIST's benchmark ASR tests: past, present, and future, *Proceedings Workshop Automatic Speech Recognition and Understanding*. pp. 483-488.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication 45*, pp. 89-95.

Polzin, T.S., Waibel, A.H. (1998). Pronunciation variations in emotional speech. In: H. Strik, J.M. Kessens, M. Wester (Eds): *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc Kerkrade, The Netherlands, pp. 103-108.

Price, P., Fisher, W., Bernstein, J., Pallet, D. (1988). The DARPA 1000-word resource management database for continuous Speech Recognition. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, pp. 651-654.

Pye, C., Wilcox, K., Siren, K. (1988). Refining transcriptions: the significance of transcriber 'errors'. *Journal of Child Language 15*, pp. 17-37.

Raymond, W.D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., Hilts, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. *Proceedings of International Conference on Spoken Language Processing (ICSLP'02)*, Denver, Colorado, USA, pp. 1125-1128.

Riley, M., Ljolje, A. (1996). Automatic generation of detailed pronunciation lexicons. In: C.-H. Lee, F. Soong, K. Paliwal (Eds.): *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, pp. 285-302.

Riley, M., Byrene, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G. (1999) Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication 29*, pp. 209-224.

Rudnicky, A.I., Hauptmann, A.G., Lee, K.-F. (1994). Survey of current speech technology. *Communications of the ACM, Vol 37, no 3*, pp. 52-57.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2001) Multiword expressions: A pain in the neck for NLP. *LinGO Working Paper (2001-03).* [http://lingo.stanford.edu/pubs/WP2001-03.ps.gz.]

Saraçlar, M., Khudanpur, S. (2000). Pronunciation Ambiguity vs Pronunciation Variability in Speech Recognition. *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'00)*, Istanbul, Turkey, pp. 1679-1682.

Scharenborg, O., Boves, L. (2002) Pronunciation Variation Modelling in a Model of Human Word Recognition. *Proceedings of Workshop on Pronunciation Modeling and Lexicon Adaptation*, Estes Park, USA, pp. 65-70.

Scherer, K.R., Giles, H. (1979). *Social Markers in Speech*. Cambridge University Press, Cambridge.

Schiel, F., Kipp, A., Tillmann, H.G., (1998). Statistical modeling of pronunciation: it's not the model, it's the data. *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* Rolduc Kerkrade, The Netherlands, pp. 131–136.

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Proceedings of International Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, USA, pp. 607-610.

Shriberg, L.D., Kwiatkowski, J., Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27*, pp. 456-465.

Shriberg, L.D., Lof, L. (1991) Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, *5*, pp. 225-279.

Slis, I.H. (1985). *The voiced-voiceless distinction and assimilation of voice in Dutch*. PhD thesis, Katholieke Universiteit Nijmegen.

Sloboda, T., Waibel, A. (1996) Dictionary Learning for Spontaneous Speech Recognition. *Proceedings of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA. pp. 2328-2331.

Smits, R., Warner, N., McQueen, J.M., Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustic Society of America*, *113*, pp. 563-574.

Strik, H., Russel, A., van den Heuvel, H., Cucchiarini, C., Boves, L. (1996) A spoken dialogue system for the Dutch public transport information service. *International Journal Speech Technology 2 (2)*, pp. 119-129.

Strik, H., Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication 29*, pp. 225-246.

Strik, H., Cucchiarini, C., Elffers, A., Binnenpoorte, D. (submitted). Multiword expressions in real-life speech: How to identify deviant pronunciation patterns. Submitted to *Phonetica* (special issue).

Swerts, M., Kloots, H., Gillis, S., De Schutter, G. (2003). Vowel reduction in spontaneous spoken Dutch. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, pp. 31-43.

Tinsley, H., Weiss, D. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *Vol. 22 (4)*, pp. 358-376.

Truong, K., Neri, A., Cucchiarini, C., Strik, H. (2004). Automatic pronunciation error detection: an acoustic-phonetic approach. *Proceedings of the InSTIL/ICALL Symposium*, Venice, Italy, pp. 135-138.

Van Bael, C., Heuvel, H. van den, Strik, H. (2005). Validation of phonetic transcriptions in the context of automatic speech recognition. Submitted to *Language Resources and Evaluation*.

Van de Velde, H. (1996). *Variatie en verandering in het gesproken Standaard-Nederlands (1935-1993)*. PhD thesis, Katholieke Universiteit Nijmegen.

Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgras, A. (1996). Effect of speaking style on LVCSR performance. *Proceedings of The Fourth International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, Addendum, pp. 16-19.

Wells, J. (1996). Why phonetic transcription is important. *Malsori - Journal of the Phonetic Society of Korea, No. 31*, pp. 239-242.

Wells, J. (2004). SAMPA computer readable phonetic alphabet. [http://www.phon.ucl.ac.uk/home/sampa/dutch.htm]

Wesenick, M-B., Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proceedings of International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA, pp. 129-132.

Wester, M., Kessens, J.M., Cucchiarini, C., Strik, H. (2001). Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer. *Language and Speech 44 (3)*, pp. 377-403.

Wester, M. (2002). *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*. PhD thesis, Katholieke Universiteit Nijmegen.

Wester, M. (2003) Pronunciation modeling for ASR - knowledge-based and data-derived methods. *Computer Speech and Language 17*, pp. 69-85.

Wissing, D., Martens, J-P., Janke, U., Goedertier, W. (2004). A spoken Afrikaans language resource designed for research on pronunciation variations. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, pp. 669-672.

Wong-Fillmore, L. (1979) Individual Differences in Second Language Acquisition. C. Fillmore, D. Kempler & W. Wang (Eds.) *Individual Differences in Language Ability and Language Behaviour*. Academic Press, New York. pp.203-228.

Woodland, P. (1998). Speech recognition. *IEEE Colloquium on Speech and Language Engineering - State of the Art (Ref. No. 1998/499)*, pp. 2/1-2/5.

Wouden, T. van der, Hoekstra, H., Moortgat, M., Renmans, B., Schuurman, I. (2002). Syntactic analysis in the Spoken Dutch Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, Spain, pp. 768-773.

Yang, Q., Martens, J.P. (2000a). Data-driven lexical modeling of pronunciation variations for ASR. *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*, Beijing, China, pp. 417-420.

Yang, Q., Martens, J.P. (2000b) On the importance of exception and cross-word rules for the data-driven creation of Lexica for ASR. *Proceedings 11th ProRisc Workshop*, Veldhoven, The Netherlands. pp. 589-593.

APPENDICES

A1 Articulatory costs matrix of CGN consonant symbols and word boundary (|), used in Align.

symbol	place	voice	nasal	stop	glide	lat	fric	trill	asp	dental	strength	del/ins
/p/	5.0	1.0	0.0	0.5	0.0	0.0	0.0	0.0	1	1	1	3.0
/b/	5.0	2.0	0.0	0.5	0.0	0.0	0.0	0.0	1	1	1	3.0
/t/	4.0	1.0	0.0	0.5	0.0	0.0	0.0	0.0	1	1	1	3.0
/d/	4.0	2.0	0.0	0.5	0.0	0.0	0.0	0.0	1	1	1	3.0
/k/	2.0	1.0	0.0	0.5	0.0	0.0	0.0	0.0	1	1	1	3.0
/g/	2.0	2.0	0.0	0.5	0.0	0.0	0.0	0.0	1	1	1	3.0
/f/	5.0	1.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/v/	5.0	2.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/s/	4.0	1.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/z/	4.0	2.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/S/	3.0	1.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/Z/	3.0	2.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/G/	2.0	2.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/x/	2.0	1.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
/m/	5.0	2.0	0.5	0.0	0.0	0.0	0.0	0.0	1	1	1	3.0
/n/	4.0	2.0	0.5	0.0	0.0	0.0	0.0	0.0	1	1	1	3.0
/N/	2.0	2.0	0.5	0.0	0.0	0.0	0.0	0.0	1	1	1	3.0
/J/	3.0	2.0	0.5	0.0	0.5	0.0	0.0	0.0	1	1	1	3.0
/l/	4.0	2.0	0.0	0.0	0.0	0.5	0.0	0.0	1	1	1	3.0
/r/	3.0	2.0	0.0	0.0	0.0	0.0	0.0	0.5	1	1	1	3.0
/w/	5.0	2.0	0.0	0.0	0.5	0.0	0.0	0.0	1	1	1	3.0
/j/	3.0	2.0	0.0	0.0	0.5	0.0	0.0	0.0	1	1	1	3.0
/h/	1.0	2.0	0.0	0.0	0.0	0.0	0.5	0.0	1	1	1	3.0
|	8.0	8.0	3.0	3.0	3.0	3.0	3.0	3.0	2	2	2	6.0

A2 Articulatory costs matrix of CGN vowel symbols and word boundary (|), used in Align.

symbol	length	front/back	tongue	round	diph	del / ins
/i/	1.5	3.0	4.0	1.0	1.0	3.0
/I/	1.0	2.5	3.5	1.0	1.0	3.0
/e/	2.0	3.0	3.0	1.0	1.5	3.0
/2/	2.0	3.0	3.0	2.0	1.5	3.0
/E/	1.0	3.0	2.0	1.0	1.0	3.0
/a/	2.0	2.0	1.0	1.5	1.0	3.0
/A/	1.0	1.0	1.5	1.5	1.0	3.0
/o/	2.0	1.0	3.0	2.0	1.5	3.0
/O/	1.0	1.0	2.0	2.0	1.0	3.0
/u/	1.5	1.0	4.0	2.0	1.0	3.0
/y/	1.5	3.0	4.0	2.0	1.0	3.0
/Y	1.0	2.5	3.5	2.0	1.0	3.0
/@/	1.0	2.0	2.5	1.5	1.0	3.0
/E+/	2.0	3.0	3.0	1.0	2.0	3.0
/Y+/	2.0	2.5	3.0	1.0	2.0	3.0
/A+/	2.0	1.5	3.0	2.0	2.0	3.0
/E:/	2.0	3.0	2.0	1.0	1.0	3.0
/Y:/	2.0	2.0	2.0	1.5	1.0	3.0
/O:/	2.0	1.0	2.0	2.0	1.0	3.0
/E~/	2.0	3.0	2.0	1.0	1.0	3.0
/A~/	2.0	1.0	1.5	1.5	1.0	3.0
/O~/	2.0	1.0	2.0	2.0	1.0	3.0
/Y~/	2.0	2.0	2.0	1.5	1.0	3.0
|	5.0	5.0	7.0	5.0	3.0	6.0

A3 Results after comparing the canonical transcription with the actual pronunciations found in spontaneous speech of highly frequent 3-grams.

gram	freq	% disagr	% sub	% del	% ins	gram	freq	% disagr	% sub	% del	% ins
ik weet niet	118	27,22	8,26	18,96	0,00	zo is dat	10	41,43	31,43	4,29	5,71
ik denk dat	113	29,40	19,76	9,64	0,00	op gegeven moment	10	39,29	7,14	32,14	0,00
in ieder geval	79	37,17	7,59	29,57	0,00	ik denk van	10	35,56	28,89	6,67	0,00
'k weet niet	75	26,86	4,76	20,76	1,33	dat moet ik	10	32,50	20,00	12,50	0,00
weet ik niet	68	29,23	22,43	6,62	0,18	over 't algemeen	10	20,00	6,92	13,08	0,00
nog een keer	59	18,64	11,23	7,42	0,00	nog wel wat	10	18,89	16,67	2,22	0,00
weet ik veel	54	26,85	24,54	2,31	0,00	dat je gewoon	10	16,00	6,00	10,00	0,00
ik ook niet	42	24,83	14,63	10,20	0,00	even kijken wat	10	10,00	5,00	2,00	3,00
af en toe	35	34,76	22,86	11,90	0,00	in dit geval	10	6,00	6,00	0,00	0,00
dat denk ik	35	22,22	18,41	3,81	0,00	dat kan natuurlijk	9	36,51	13,49	23,02	0,00
't is gewoon	35	19,68	2,86	16,83	0,00	tot en met	9	36,11	20,83	15,28	0,00
op zich wel	35	18,93	16,79	2,14	0,00	dat zou wel	9	34,57	17,28	17,28	0,00
'ns een keer	34	31,51	11,76	18,91	0,84	dat is eigenlijk	9	34,34	17,17	17,17	0,00
dat is gewoon	30	19,00	8,33	10,67	0,00	lijkt me ook	9	30,56	11,11	19,44	0,00
dat zal wel	27	25,10	16,05	9,05	0,00	ook helemaal niet	9	27,78	6,48	21,30	0,00
wat is dat	24	26,04	18,23	7,29	0,52	dat hoeft niet	9	26,67	8,89	17,78	0,00
dat is echt	24	21,88	15,10	6,77	0,00	even kijken of	9	23,46	13,58	9,88	0,00
hoe heet dat	23	30,43	23,37	6,52	0,54	dat was eigenlijk	9	18,52	7,41	11,11	0,00
wat dat betreft	23	20,07	9,70	10,37	0,00	ik hoop dat	9	18,06	11,11	6,94	0,00
hebben we nog	22	20,71	2,02	18,69	0,00	in eerste instantie	9	17,78	2,96	14,81	0,00
wat zeg je	21	20,83	16,67	4,17	0,00	nog steeds niet	9	13,13	5,05	8,08	0,00
voor de rest	20	20,56	10,00	10,56	0,00	zijn we weer	9	11,11	4,17	6,94	0,00
de vorige keer	19	31,58	12,92	18,66	0,00	kan ook nog	9	9,72	6,94	2,78	0,00
maar goed dat	19	30,99	22,22	8,77	0,00	wat wil je	9	4,17	1,39	2,78	0,00
in elk geval	18	23,33	11,11	12,22	0,00	eigenlijk niet zo	8	47,73	17,05	29,55	1,14
't is natuurlijk	17	45,59	11,76	33,82	0,00	volgens mij niet	8	37,50	12,50	25,00	0,00
weet 'k niet	17	38,66	15,13	21,01	2,52	gezegd van nou	8	32,95	17,05	15,91	0,00
dat kan niet	17	26,80	3,92	22,22	0,65	dat zei ik	8	32,14	19,64	5,36	7,14
dat dacht ik	17	20,92	20,26	0,65	0,00	dat is inderdaad	8	27,88	12,50	15,38	0,00
maar volgens mij	16	41,15	11,46	29,17	0,52	dat zou kunnen	8	23,75	12,50	10,00	1,25
volgens mij wel	16	35,42	11,46	23,96	0,00	ik weet het	8	23,21	23,21	0,00	0,00
dat is natuurlijk	16	34,62	15,38	19,23	0,00	wat zei je	8	23,21	16,07	7,14	0,00
op die manier	16	31,94	25,00	6,94	0,00	ik wist niet	8	22,22	9,72	12,50	0,00
op dit moment	16	25,57	19,89	5,68	0,00	kan ik niet	8	21,88	14,06	7,81	0,00
gewoon een beetje	16	21,35	9,90	11,46	0,00	kan ook niet	8	20,31	12,50	7,81	0,00
dat bedoel ik	16	21,25	10,00	10,63	0,63	ik geloof dat	8	17,50	13,75	3,75	0,00
dat kan ook	16	17,19	6,25	10,94	0,00	hebben we niet	8	16,67	0,00	16,67	0,00
ik denk nou	15	27,50	20,83	6,67	0,00	denk ik dat	8	13,89	9,72	4,17	0,00
dat wil ik	15	20,83	14,17	6,67	0,00	heel erg veel	8	12,50	11,11	0,00	1,39
dat snap ik	14	22,22	15,08	7,14	0,00	kan ik wel	8	10,94	9,38	1,56	0,00
een soort van	13	18,80	14,53	4,27	0,00	wel 'ns wat	8	10,94	9,38	1,56	0,00
dat kan wel	13	14,53	11,11	3,42	0,00	tot nu toe	8	10,71	10,71	0,00	0,00
min of meer	13	11,54	10,58	0,96	0,00	je weet wel	8	9,38	7,81	1,56	0,00
volgens mij ook	12	40,15	15,91	21,97	2,27	wel of niet	8	9,38	6,25	3,13	0,00
als 't ware	12	23,15	0,93	22,22	0,00	je moet wel	8	7,81	4,69	3,13	0,00

denk 't niet	12	21,30	11,11	10,19	0,00	dat mag wel	7	28,57	19,05	9,52	0,00
'ns even kijken	12	17,59	9,26	5,56	2,78	ook niet helemaal	7	26,19	14,29	11,90	0,00
doe je dat	12	8,33	4,76	3,57	0,00	geloof ik of	7	25,40	19,05	6,35	0,00
dank je wel	12	0,93	0,00	0,93	0,00	dat gaat wel	7	22,22	14,29	7,94	0,00
zoiets van ja	11	57,27	32,73	18,18	6,36	lijkt me toch	7	19,05	7,94	11,11	0,00
nou volgens mij	11	38,84	14,88	23,97	0,00	heen en weer	7	17,86	7,14	10,71	0,00
is natuurlijk ook	11	35,61	11,36	23,48	0,76	dat mag niet	7	17,46	7,94	7,94	1,59
zo van ja	11	29,87	20,78	9,09	0,00	niet zo lekker	7	15,71	10,00	5,71	0,00
ik denk niet	11	28,28	18,18	10,10	0,00	nog een keertje	7	14,29	3,90	10,39	0,00
't is allemaal	10	43,00	16,00	27,00	0,00	heel lang geleden	7	11,90	3,57	8,33	0,00

A4 Results after comparing the canonical transcription with the actual pronunciations found in spontaneous speech of highly frequent 4-grams.

gram	freq	% disagr	% sub	% del	% ins
op een gegeven moment	50	47,13	7,38	39,63	0,13
dat weet ik niet	31	29,03	19,65	9,38	0,00
ja dat weet ik	19	20,53	15,26	5,26	0,00
dat denk ik ook	19	18,66	6,70	11,96	0,00
als 't goed is	14	18,57	3,57	15,00	0,00
'k weet niet of	14	29,37	5,56	21,43	2,38
aan de andere kant	14	8,16	1,53	6,63	0,00
of wat dan ook	13	31,54	20,00	11,54	0,00
denk ik ook wel	12	17,42	12,12	5,30	0,00
dat maakt niet uit	11	42,42	14,39	20,45	7,58
weet ik veel wat	11	26,45	20,66	5,79	0,00
dat weet ik nog	10	24,55	20,91	3,64	0,00
dat vind 'k ook	9	48,89	31,11	17,78	0,00
dat weet 'k niet	9	42,22	17,78	23,33	1,11
ik denk niet dat	9	31,48	15,74	15,74	0,00
maakt toch niet uit	8	31,25	5,21	20,83	5,21
dat denk ik wel	8	18,75	7,29	11,46	0,00
op een andere manier	8	10,83	8,33	2,50	0,00
dat is niet zo	7	40,00	25,71	14,29	0,00
'k weet niet precies	7	28,57	5,49	23,08	0,00
dat vind ik wel	7	30,95	17,86	13,10	0,00

A5 Results after alignment of individual words both in MWE context and in any other context in 3-grams.

		in MWE context					in other context				
word	is part of n-gram	freq	%sub	%del	%ins	%dis	freq	%sub	%del	%ins	%dis
‘t	’t is natuurlijk	17	14,71	35,29	0,00	50,00	4843	8,96	21,85	0,47	31,28
is	’t is natuurlijk	17	14,71	11,76	0,00	26,47	4190	16,34	9,92	1,86	28,11
natuurlijk	’t is natuurlijk	17	10,29	38,97	0,00	49,26	476	11,74	25,45	0,11	37,29
af	af en toe	35	20,00	0,00	0,00	20,00	86	6,40	2,91	0,00	9,30
en	af en toe	35	47,14	35,71	0,00	82,86	4291	6,90	7,04	1,60	15,53
toe	af en toe	35	1,43	0,00	0,00	1,43	75	0,00	0,00	0,00	0,00
als	als 't ware	12	0,00	30,56	0,00	30,56	1349	8,03	39,39	0,02	47,44
‘t	als 't ware	12	4,17	45,83	0,00	50,00	5285	9,32	20,91	1,24	31,47
ware	als 't ware	12	0,00	4,17	0,00	4,17	1	0,00	0,00	0,00	0,00
dat	dat is natuurlijk	16	12,50	4,17	0,00	16,67	7248	18,02	9,42	0,03	27,47
is	dat is natuurlijk	16	28,13	6,25	0,00	34,38	4191	16,28	9,94	1,86	28,08
natuurlijk	dat is natuurlijk	16	13,28	28,13	0,00	41,41	477	11,64	25,84	0,10	37,58
hoe	hoe heet dat	23	4,35	10,87	2,17	17,39	617	0,57	2,92	0,00	3,48
heet	hoe heet dat	23	18,84	7,25	0,00	26,09	77	9,96	3,03	0,43	13,42
dat	hoe heet dat	23	40,58	2,90	0,00	43,48	6086	18,28	8,24	0,02	26,54
in	in ieder geval	79	10,13	12,03	0,00	22,15	2152	5,37	6,02	0,07	11,45
ieder	in ieder geval	79	0,95	48,73	0,00	49,68	20	1,25	13,75	0,00	15,00
geval	in ieder geval	79	11,90	21,27	0,00	33,16	45	14,67	12,00	0,00	26,67
ook	ook helemaal niet	9	0,00	0,00	0,00	0,00	2860	17,67	0,93	0,03	18,64
helemaal	ook helemaal niet	9	11,11	26,98	0,00	38,10	463	9,41	25,46	0,00	34,87
niet	ook helemaal niet	9	0,00	22,22	0,00	22,22	151	1,32	20,97	0,00	22,30
op	op die manier	16	40,63	15,63	0,00	56,25	1491	12,68	1,81	0,10	14,59
die	op die manier	16	18,75	3,13	0,00	21,88	3855	9,26	2,78	0,08	12,11
manier	op die manier	16	21,25	5,00	0,00	26,25	42	16,67	5,71	0,00	22,38
weet	weet ik niet	68	26,47	1,47	0,00	27,94	945	10,02	6,81	0,00	16,83
ik	weet ik niet	68	46,32	2,21	0,74	49,26	6332	19,24	5,54	0,77	25,55
niet	weet ik niet	68	2,45	14,71	0,00	17,16	3876	2,92	15,33	0,03	18,27
zoiets	zoiets van ja	11	25,45	18,18	12,73	56,36	99	19,19	8,89	11,52	39,60
van	zoiets van ja	11	39,39	30,30	0,00	69,70	2504	30,56	6,27	0,00	36,83
ja	zoiets van ja	11	40,91	0,00	0,00	40,91	8406	4,39	0,32	0,92	5,63

A6 Results after alignment of individual words both in MWE context and in any other context in 4-grams.

		in MWE context					in other context				
word	is part of n-gram	freq	%sub	%del	%ins	%dis	freq	%sub	%del	%ins	%dis
dat	dat is niet zo	7	23,81	19,05	0,00	42,86	5398	16,51	13,10	0,02	29,63
is	dat is niet zo	7	35,71	7,14	0,00	42,86	3160	17,33	10,16	1,72	29,21
niet	dat is niet zo	7	0,00	23,81	0,00	23,81	2860	2,17	19,85	0,02	22,04
zo	dat is niet zo	7	57,14	0,00	0,00	57,14	1600	37,75	0,56	0,09	38,41
dat	dat maakt niet uit	11	30,30	18,18	0,00	48,48	5394	16,49	13,10	0,02	29,61
maakt	dat maakt niet uit	11	20,45	22,73	0,00	43,18	57	5,70	17,11	0,00	22,81
niet	dat maakt niet uit	11	0,00	33,33	0,00	33,33	2853	2,17	19,82	0,02	22,01
uit	dat maakt niet uit	11	0,00	0,00	45,45	45,45	400	6,00	0,88	5,38	12,25
dat	dat vind 'k ook	9	11,11	22,22	0,00	33,33	5396	16,53	13,10	0,02	29,65
vind	dat vind 'k ook	9	47,22	27,78	0,00	75,00	346	29,99	25,29	0,00	55,27
‘k	dat vind 'k ook	9	22,22	0,00	0,00	22,22	776	23,07	0,77	10,95	34,79
ook	dat vind 'k ook	9	33,33	0,00	0,00	33,33	2545	15,83	1,24	1,83	18,90
dat	dat weet ik niet	31	15,05	21,51	0,00	36,56	5374	16,53	13,06	0,02	29,61
weet	dat weet ik niet	31	27,96	1,08	0,00	29,03	776	9,92	8,81	0,00	18,73
ik	dat weet ik niet	31	38,71	3,23	0,00	41,94	3727	21,34	7,04	0,74	29,13
niet	dat weet ik niet	31	3,23	9,68	0,00	12,90	2979	2,92	15,64	0,03	18,60
dat	dat weet ik nog	10	23,33	10,00	0,00	33,33	5395	16,51	13,12	0,02	29,64
weet	dat weet ik nog	10	20,00	0,00	0,00	20,00	797	10,50	8,62	0,00	19,11
ik	dat weet ik nog	10	45,00	0,00	0,00	45,00	3748	21,42	7,03	0,73	29,19
nog	dat weet ik nog	10	3,33	3,33	0,00	6,67	1482	8,50	3,55	0,02	12,08
of	of wat dan ook	13	23,08	7,69	0,00	30,77	854	15,93	3,57	0,12	19,61
wat	of wat dan ook	13	0,00	33,33	0,00	33,33	1001	8,33	6,93	0,03	15,28
dan	of wat dan ook	13	48,72	0,00	0,00	48,72	2133	15,27	13,11	0,02	28,40
ook	of wat dan ook	13	3,85	0,00	0,00	3,85	2541	15,96	1,24	1,83	19,03
op	op een gegeven moment	50	21,00	6,00	0,00	27,00	848	12,56	3,77	0,18	16,51
een	op een gegeven moment	50	2,00	47,00	1,00	50,00	3124	14,37	15,51	1,47	31,35
gegeven	op een gegeven moment	50	3,00	51,67	0,00	54,67	20	6,67	27,50	0,00	34,17
moment	op een gegeven moment	50	9,00	36,33	0,00	45,33	50	7,67	10,33	0,00	18,00
‘k	‘k weet niet precies	7	14,29	0,00	0,00	14,29	1426	15,85	0,63	9,26	25,74
weet	k weet niet precies	7	4,76	28,57	0,00	33,33	1131	9,99	11,64	0,00	21,63
niet	k weet niet precies	7	0,00	23,81	0,00	23,81	3601	1,96	20,47	0,02	22,45
precies	k weet niet precies	7	7,14	23,81	0,00	30,95	151	4,30	19,98	1,21	25,50
als	als 't goed is	14	4,76	26,19	0,00	30,95	1413	9,74	36,73	0,02	46,50
‘t	als 't goed is	14	0,00	35,71	0,00	35,71	4294	9,55	23,44	1,30	34,29
goed	als 't goed is	14	2,38	0,00	0,00	2,38	599	7,29	4,34	0,00	11,63
is	als 't goed is	14	7,14	0,00	0,00	7,14	2043	15,32	5,43	2,20	22,96
weet	weet ik veel wat	11	27,27	0,00	0,00	27,27	1050	14,29	7,21	0,00	21,49
ik	weet ik veel wat	11	0,00	0,00	0,00	0,00	5217	21,04	7,21	0,81	29,06
veel	weet ik veel wat	11	48,48	21,21	0,00	69,70	492	30,62	5,08	0,14	35,84
wat	weet ik veel wat	11	0,00	0,00	0,00	0,00	1113	8,60	2,64	0,03	11,26

SUMMARIES

ENGLISH AND DUTCH

Summary

Chapter 1: Introduction

Everyday spoken communication between language users is based on the ability of speakers to produce intelligible speech, and, at the same time, on the ability of listeners to understand the spoken message. The latter is remarkable given the fact that each time a word is uttered, its pronunciation can differ. A quick analysis of the phonetic transcriptions of a small sample of speech, recorded in a real-life situation (see page 3) already shows that the pronunciation of the same words uttered by one and the same speaker can differ each time a word is uttered and can be rather different from the dictionary or canonical pronunciation. For research on pronunciation phenomena numerous samples of real-life speech need to be collected. If properly annotated, such collections of speech are referred to as *speech corpora*. Speech corpora constitute an essential resource for empirical investigations on spoken language for a broad range of researchers. Usually large speech corpora only contain an orthographic transcription of the speech material, whereas to be useful as resource for pronunciation research, speech corpora should also contain *phonetic transcriptions* of the speech. The research reported in this thesis is focused on, first, the generation of phonetic transcriptions of large speech corpora; second, and in relation to this, gathering new phonological knowledge of less researched speech styles, and third, the evaluation of the quality of the phonetic transcriptions.

Generating phonetic transcriptions of sufficient quality for large amounts of real-life speech is not a straightforward task. First, manual phonetic transcriptions are difficult to make, are time-consuming and thus expensive, and contain an element of subjectivity. Because of the sheer size of present-day speech corpora, a complete manual phonetic transcription made by an expert phonetician is practically impossible, given the constraints on budget and time. Therefore, alternative transcription procedures, such as (semi-) automatic procedures, need to be developed. These (semi-) automatic techniques have a substantial cost advantage over manually generated transcriptions, while they are more consistent and certainly more reproducible than transcriptions of an individual expert.

Second, the pronunciation of words in real-life speech, especially spontaneous speech, deviates to a large extent from the canonical representations found in pronunciation dictionaries. So, a one-to-one translation from orthographic words to their canonical representation found in a dictionary is unlikely to result in an accurate representation of the speech signal. In order to model pronunciation phenomena to obtain better phonetic transcriptions of speech styles that do not resemble the canonical representation, knowledge, either pronunciation rules or statistical information, needs to be collected.

Since most of the speech research and literature is based on read speech, which is typically pronounced more carefully than spontaneous speech, methods need to be developed to obtain phonological knowledge of spontaneous speech. One source of knowledge is formed by already existing (medium-sized) speech corpora that contain high quality phonetic transcription of speech styles similar to the type speech that needs to be transcribed. In case such a corpus is indeed available, the challenge is how to extract and represent the phonological knowledge. For instance, it still needs to be considered to what extent derived pronunciation rules are able to capture the variety of pronunciation in spontaneous speech. In other circumstances, if one is working on a language for which no corpora are yet compiled, or if the speech to be transcribed does not resemble any of the speech styles in already existing corpora, the problem of obtaining knowledge becomes more challenging. In this thesis automatic phonetic transcription procedures are proposed and tested for both circumstances.

Third, the development of automatic transcription procedures and the deployment of new phonological knowledge to improve automatic phonetic transcription procedures can only be considered as useful if the resulting transcriptions are of sufficient quality. Automatic phonetic transcriptions can only replace human-made transcriptions if their quality is good enough. In this thesis considerable attention is paid to the evaluation of phonetic transcriptions. For both an automatic and a human-made transcription the intention is to represent the actual pronunciation. Consequently, in order to measure transcription quality it is necessary to determine to what extent the transcriptions deviate from the actual pronunciation. This can be accomplished by comparing the transcriptions to a reference transcription that represents the ground truth. However, since there is no such thing as an absolute true transcription – transcriptions always suffer from subjectivities, even if made by an expert – such a reference transcription can only be approximated. In this thesis the reference transcription was defined by a consensus transcription. Comparing transcriptions with this consensus transcription yields a more objective quality measure. In this way the quality of both automatically generated and human-made transcription was measured and compared to other measures.

The experiments described in this thesis were all carried out on the Spoken Dutch Corpus, CGN (Corpus Gesproken Nederlands), containing 9 million words of contemporary Dutch as spoken in Flanders and the Netherlands. Parts of this thesis describe the generation of the automatic phonetic transcriptions for the Northern Dutch part of the corpus. For about one million words a more detailed annotation is available, i.e. manually verified broad phonetic transcriptions. These manually verified transcriptions are also subject of investigation in this thesis.

In addition to the introductory chapter, in which the basic issues related to phonetic transcriptions of large speech corpora are introduced, this thesis consists of four chapters (chapters 2 to 5) describing the experiments that were conducted to reach the goals presented above. Chapters 2 and 3 are concerned with both automatic phonetic transcription generation methods and measures to establish transcription quality. Chapter 4 focuses on manual phonetic transcription generation and evaluation methods. Chapter 5 describes an exploratory study on a specific phenomenon in spontaneous speech. In the concluding chapter (chapter 6) the main conclusions are presented based on the results of the experiments.

Chapter 2: A procedure for the production of phonetic transcriptions of large speech corpora

In this chapter an automatic phonetic transcription procedure is proposed that can circumvent the problem of lacking knowledge about phonological processes. The reason for developing this procedure was both the absence of well-described phonological rules in the literature, and the absence of phonetically transcribed speech corpora containing speech styles similar to the speech to be transcribed. We propose an iterative procedure that is intended to improve the efficiency of generating transcriptions for large speech corpora, and at the same time to obtain new knowledge with respect to the nature and frequency of phonological processes in various speech styles. This dual-purpose procedure consists of several cycles, where in each cycle an automatic transcription is generated for a sample of the corpus. From one cycle to the other the transcription generation techniques become more complex, starting from a simple lexicon look-up, to a generation method in which an Automatic Speech Recognition (ASR) system is used. After each cycle the automatic transcription is compared to a reference transcription to establish the quality of the automatic transcription. In order to determine whether the automatically generated transcription is good enough, its quality is compared to a threshold that is set on the basis of human-made transcriptions obtained in similar tasks. In determining the threshold the uncertainties of human-made transcription quality are taken into account. Once the threshold is reached for a certain speech style in one of the cycles, subsequent cycles with more complicated transcription techniques can be omitted. Besides continuous quality assessments in each cycle, a detailed analysis is performed to obtain information that can be used to improve the automatic transcription for the subsequent cycle.

The procedure is demonstrated with real-life data from the CGN. The experiments show that the quality of the automatic transcriptions is improved by applying the cycles of the procedure, and that for read speech similar quality levels can be reached as with

human-made transcriptions. The results also reveal that the procedure succeeded in providing new and more complete information on the nature and frequency of various phonological processes in different speech styles. It can be concluded that the dual-purpose technique appeared to be effective both for obtaining automatic transcriptions of good quality and for extracting new systematic phonological knowledge from unexplored speech data. In the specific case of Dutch very useful information was obtained on the frequency of application of the various processes in the different speech styles, whereas the nature of these processes already signalled in the literature for read speech were confirmed.

Chapter 3: Variant-based pronunciation variation modelling for automatic phonetic transcription of spontaneous speech

The goal of the experiments reported on in this chapter was to improve automatic phonetic transcriptions of spontaneous speech. To reach that goal an attempt was made to improve the modelling of pronunciation variation. Often, pronunciation variants are generated by rules and then listed in a lexicon. This lexicon is input for an ASR system that is used in forced recognition mode: the ASR system chooses that variant from the lexicon that best matches the speech signal. In this way an automatic phonetic transcription can be obtained. In spontaneous speech, pronunciation variation can be rather extreme in the sense that it is difficult to capture the rich variation in rewrite rules that are subsequently used to generate plausible pronunciation variants for the lexicon. In this chapter an alternative method is tested for obtaining pronunciation variants. We used a manually transcribed corpus of spontaneous Dutch (part of the core corpus of the CGN) as the source for pronunciation variants. The most frequent variants were extracted from the corpus and stored in the lexicon. Information on the prior probabilities of these variants was also captured and used by the ASR system while choosing the best matching variant. The automatic transcription generated by this variant-based pronunciation variation modelling showed an improvement of 6% (20% relative) over a lexicon look-up procedure. Although the improvement is substantial, still the quality level of human-made transcriptions for spontaneous speech cannot be reached.

From the results it can be concluded that the adopted method is indeed suitable for improving automatic transcription of spontaneous speech and that a large amount of manually transcribed phonetic data is an extremely useful source for collecting pronunciation variants and their prior probabilities. It is clear that the more transcribed data are available, the better spontaneous speech can be modelled, which, in turn, means that automatic phonetic transcription can be improved such that more transcriptions can become available at lower costs.

Chapter 4: Measuring phonetic transcription quality in large speech corpora

In order to reduce transcription time and therefore costs human-made transcriptions in large speech corpora projects are usually produced by following a procedure in which several transcribers edit an example transcription. The quality of this type of transcriptions is usually established by measuring inter-transcriber agreement. We argue that this is not a suitable measure, since first, similarity between phonetic symbols does not necessarily mean that the symbol represents the actual pronunciation, and second, percentages of agreement may be artificially inflated because of the bias effects of the example transcription. Therefore, an additional measure is introduced to establish transcription quality, i.e. the distance between the transcription and a consensus transcription which can be considered as a close approximation of a true reference transcription.

An experiment was set up in which several individual transcribers transcribe the same speech sample by editing an example transcription. At the same time two expert transcribers were asked to make the consensus transcription from scratch. Both inter-transcriber agreement as well as agreement between the consensus transcription and each of the individual transcriptions was measured.

The results show that the individual transcribers do indeed improve the given example transcription. However, the average percentages of deviations between the individual transcriptions (inter-transcriber disagreement) are, in most speech styles, significantly smaller than the average distance between the individual transcriptions and the consensus transcription. This provides evidence for the assumption that the example transcription artificially inflated the inter-transcriber agreement scores. Furthermore, it was clear that the added value of human transcribers editing an example transcription was not equally large for all speech styles. Although human transcribers significantly improved the example transcription of read speech, still the transcribers systematically confused voiced and unvoiced phonemes. Therefore, we suggest that future projects on transcriptions of large amounts of speech data should think twice before hiring expensive human transcribers for phonetic transcriptions of read speech. Finally, it is demonstrated that if transcriptions are made by editing example transcriptions, human transcriptions, of any type or quantity, should not be evaluated by solely establishing an inter-transcriber agreement score, but more objective measures, such as a comparison with a consensus transcription, should be taken into account as well.

Chapter 5: Multiword Expressions in spoken language

The availability of a large corpus of phonetically transcribed spontaneous speech not only offers the possibility of extracting frequent pronunciation variants as demonstrated in

chapter 3. Additionally, large phonetically transcribed corpora offer the opportunity of analysing pronunciation phenomena on a higher level, i.e., beyond word boundaries. The question chapter 5 aims to answer is: are there systematic patterns that can predict the occurrence of some of the extreme pronunciations that occur in spontaneous speech? This type of information can be used for better modelling spontaneous speech processes when generating automatic phonetic transcriptions.

For the experiment, an inventory was drawn up of frequently found word sequences (N-grams) that were extracted from orthographic transcriptions of spontaneous speech in the CGN. For a selection of these N-grams the phonetic transcriptions were examined and we found that the pronunciation of these N-grams differed to a large extent from the canonical representations. More importantly, we found that words within the N-gram context showed different pronunciation patterns than the same words when they occur in any other context. This suggests that the N-grams that were investigated may be considered as Multiword Expressions and should be treated as one entry with their own specific pronunciation variants in the lexicons of ASR systems used for automatic phonetic transcriptions.

We can conclude that phonetically transcribed corpora are a valuable source for research into phenomena and problems that affect automatic phonetic transcriptions of spontaneous speech and that have so far been elusive. To obtain these corpora, it is essential to develop methods for generating more accurate automatic phonetic transcriptions. The results obtained with the medium size corpus that was used in this study already show a number of promising directions for that research.

Chapter 6: General discussion and conclusions

The last chapter summarizes the conclusions from the four experiments, and discusses the results against the background of related research elsewhere. Finally, several recommendations are made for future research aimed at improving the accuracy of automatic phonetic transcriptions of notoriously difficult sounds, such as the distinction between voiced and unvoiced fricatives, and the presence or absence of the schwa, /r/, /n/ and /t/.

For the central research topic in this thesis – phonetic transcriptions of large speech corpora – the large speech corpora that are now available appear to be extremely valuable in various respects. These large speech corpora allow us to study pronunciation phenomena for the benefit of improving pronunciation modelling, and for the aim of developing more refined and more adequate methods for transcription generation and transcription assessment. In this way, phonetic transcription research does not only profit from large collections of speech, but also contributes to the future development of new, large and

more comprehensive speech corpora for which phonetic transcriptions need to be generated.

Samenvatting (summary in Dutch)

Hoofdstuk 1: Inleiding

De dagelijkse communicatie tussen taalgebruikers bestaat bij de gratie van zowel het vermogen van sprekers om verstaanbaar te spreken als bij het vermogen van de luisteraar om spraak te verstaan. Dat laatste is opmerkelijk en bijzonder, gezien het feit dat de uitspraak van een woord iedere keer als het wordt uitgesproken kan verschillen. Een analyse van een handgemaakte fonetische transcriptie van slechts een kort spraakfragment uit een alledaagse conversatie (zie pagina 3), laat al zien dat een woord zelfs verschillend wordt uitgesproken door één en dezelfde spreker in twee uitingen die vlak na elkaar zijn uitgesproken. Bovendien kan de uitspraak behoorlijk afwijken van de canonieke uitspraak zoals die in uitspraakwoordenboeken is weergegeven. Om onderzoek te kunnen doen naar uitspraakverschijnselen moeten meer spraakfragmenten worden verzameld. Een dergelijke verzameling spraakfragmenten met tenminste een orthografische transcriptie wordt een spraakcorpus genoemd. Spraakcorpora vormen voor veel wetenschappers op het gebied van taal en spraak een essentiële bron voor empirisch onderzoek naar verschijnselen in gesproken taal. Echter, voordat een spraakcorpus daadwerkelijk gebruikt kan worden voor onderzoek gericht op uitspraakverschijnselen, zal de spraak in een dergelijk corpus moeten zijn voorzien van fonetische transcripties. Het onderzoek dat in dit proefschrift is beschreven heeft drie hoofddoelen. Het eerste doel is om fonetische transcripties voor grote spraakcorpora te genereren, en hiermee samenhangend, om, ten tweede, nieuwe kennis te vergaren over fonologische processen die plaatsvinden in spraakstijlen waarnaar nog niet zo veel onderzoek is gedaan. En tot slot, het derde doel is om de kwaliteit van fonetische transcripties van grote spraakcorpora te bepalen en te evalueren.

Het genereren van hoge kwaliteit fonetische transcripties van grote spraakcorpora is geen eenvoudige taak. Het eerste probleem heeft te maken met het feit dat het maken van fonetische transcripties moeilijk, tijdrovend en daarom ook kostbaar is en dat ze het resultaat zijn van de subjectieve waarneming van de transcribent. Aangezien de hoeveelheid spraak in de hedendaagse corpora enorm groot is, is het ook nog eens, gezien beperkingen van tijd en geld, onmogelijk om alle spraak te laten transcriberen door een deskundige foneticus. Een alternatieve, (semi-)automatische transcriptieprocedure zou uitkomst kunnen bieden. De voordelen van zo'n (semi-) automatische transcriptieprocedure ten opzichte van een procedure waarin transcripties met de hand gemaakt worden, zijn de lagere kosten, de hogere mate van consistentie en de reproduceerbaarheid.

Het tweede probleem dat zich voordoet, zijn de grote verschillen tussen de uitspraak van woorden in dagelijkse spraak, vooral spontane spraak, en de canonieke uitspraak zoals

die is te vinden in lexica. Zouden de woorden uit een spontaan gesproken uiting worden vervangen door de canonieke representaties, dan is het zeer onwaarschijnlijk dat die canonieke representaties een correcte weergave zijn van het spraaksignaal. Om betere fonetische transcripties te kunnen maken voor die spraakstijlen, die wat betreft uitspraak sterk afwijken van de canonieke representatie, moet de variatie die zich voordoet in de uitspraak worden gemodelleerd. Maar om uitspraakverschijnselen op een of andere manier te kunnen modelleren is wel voldoende kennis nodig. Het type kennis dat nodig is voor het modelleren van uitspraakvariatie is op dit moment niet te vinden in de literatuur aangezien er onvoldoende onderzoek is gedaan naar spontane spraakstijlen. Daarom moet er een manier gevonden worden waarop die kennis vergaard kan worden. Een mogelijke kennisbron bestaat uit reeds bestaande spraakcorpora die soortgelijke spraakfragmenten bevatten die al fonetisch getranscribeerd zijn. Zelfs als deze corpora beschikbaar zijn, moet er nog gezocht worden naar een manier om de kennis over uitspraak uit die corpora te halen en uiteindelijk te kunnen representeren. Het is de vraag of fonologische herschrijfregels toereikend zijn om de brede en rijke uitspraakvariatie van spontane spraak te beschrijven. In de gevallen dat dergelijke spraakcorpora niet beschikbaar zijn, doordat bijvoorbeeld gewerkt wordt aan een taal waarvoor nog geen corpora zijn gemaakt, of doordat gewerkt wordt met een type spraak dat in geen ander corpus voorkomt, wordt het probleem van de benodigde kennis vergaren nog groter. In dit proefschrift worden automatische fonetische transcriptieprocedures ontwikkeld en getest voor beide situaties.

Het derde probleem heeft te maken met de kwaliteit. Immers, het ontwikkelen van procedures en het inzetten van nieuw vergaarde kennis is alleen zinvol als de uiteindelijke transcripties van voldoende kwaliteit zijn. Voor een automatisch gegenereerde transcriptie geldt dat deze als goed genoeg kan worden beschouwd als de kwaliteit vergelijkbaar is met de kwaliteit van een handgemaakte transcriptie. In dit proefschrift wordt dan ook veel aandacht besteed aan de evaluatie van fonetische transcripties resulterend uit zowel automatische als handmatige procedures. De bedoeling van beide transcriptieprocedures is natuurlijk om het spraaksignaal zo goed mogelijk weer te geven. Om vast te kunnen stellen in hoeverre dit doel bereikt is, zullen de resulterende transcripties vergeleken moeten worden met een referentietranscriptie die de werkelijke uitspraak precies weergeeft. Maar gezien het feit dat een eenduidig ware referentietranscriptie niet bestaat – transcripties zullen altijd gebaseerd zijn op subjectieve oordelen, zelfs als deze gemaakt worden door een expert – moet gezocht worden naar een alternatieve transcriptie die de waarheid zo goed mogelijk benadert. In dit proefschrift is gebruik gemaakt van een consensustranscriptie die kan dienen als referentietranscriptie. Door transcripties uit een automatische en handmatige procedure te vergelijken met de consensustranscriptie kan een objectiever oordeel worden gevormd over de kwaliteit van die transcripties.

Voor de experimenten die in dit proefschrift zijn beschreven, is gebruik gemaakt van het Corpus Gesproken Nederlands (CGN). Het CGN bevat ongeveer 9 miljoen woorden hedendaags Nederlands zoals gesproken in Vlaanderen en Nederland. Zowel de automatische als de handmatige geverifieerde brede fonetische transcripties van de spraak in dit corpus vormen onderwerp van onderzoek in dit proefschrift.

Naast een inleidend hoofdstuk waarin de probleemstelling rondom fonetische transcripties is beschreven en een concluderend hoofdstuk, bestaat dit proefschrift uit vier hoofdstukken waarin de experimenten die zijn uitgevoerd worden beschreven. In hoofdstuk 2 en 3 worden methoden voor het genereren en evalueren van automatische fonetische transcripties beschreven. Hoofdstuk 4 is gewijd aan het genereren en evalueren van handmatige fonetische transcripties en in hoofdstuk 5 wordt een verkennend onderzoek beschreven over uitspraakverschijnselen in spontane spraak. In het laatste hoofdstuk, hoofdstuk 6, worden de belangrijkste conclusies gepresenteerd die getrokken kunnen worden op basis van de experimenten.

Hoofdstuk 2: Een procedure voor de productie van fonetische transcripties van grote spraakcorpora

In dit hoofdstuk is een automatische fonetische transcriptieprocedure beschreven. Het uitgangspunt bij deze procedure was een situatie waarbij onvoldoende fonologische kennis aanwezig is, wegens gebrek aan beschrijvingen in de literatuur en wegens gebrek aan soortgelijke, reeds beschikbare getranscribeerde corpora. De procedure is iteratief en heeft twee doelen, namelijk op een efficiënte manier automatische fonetische transcripties te genereren, en nieuwe kennis op te doen over de aard en het voorkomen van allerlei processen in verschillende spraakstijlen. De iteratieve procedure bestaat uit verschillende cycli, waarbij in iedere cyclus een automatische transcriptie wordt gegenereerd, geëvalueerd en geanalyseerd. Met iedere volgende cyclus wordt de methode om transcripties te genereren complexer, te beginnen met een eenvoudige concatenatie van canonieke representaties tot een methode waarin een Automatisch Spraakherkenningssysteem (ASH) is ingezet. Het voordeel van de iteratieve aanpak is dat de meer complexe technieken achterwege kunnen worden gelaten als bij de evaluatie in een initiële cyclus blijkt dat de kwaliteit van de transcriptie van een bepaalde spraakstijl al goed genoeg is. De kwaliteit wordt gemeten door een vergelijking te maken met een consensustranscriptie van hetzelfde materiaal, waarna het verkregen kwaliteitsniveau wordt vergeleken met een drempelwaarde die is vastgesteld op basis van prestaties van menselijke transcribenten voor soortgelijke spraakstijlen. Tot slot levert de analyse in

iedere cyclus nieuwe kennis op over uitspraakverschijnselen en die kennis kan direct worden ingezet om de transcriptie in de daaropvolgende cyclus te verbeteren.

Aan de hand van data uit het CGN, bestaande uit verschillende soorten spraak opgenomen in verschillende conversationele situaties, is de iteratieve procedure gedemonstreerd. De resultaten van de experimenten laten zien dat na iedere cyclus de automatische transcriptie inderdaad beter wordt en dat voor voorgelezen spraak vergelijkbare kwaliteit kan worden gehaald als met handgemaakte transcripties. Daarnaast is gebleken dat de procedure die gevolgd is, geschikt is om nieuwe kennis te verkrijgen over de aard en het voorkomen van allerlei fonologische processen in verschillende soorten spraak. Kortom, de twee doelen die met de iteratieve procedure zijn nagestreefd, betere automatische fonetische transcripties en meer kennis van uitspraakverschijnselen in nog niet intensief onderzochte spraakstijlen, zijn bereikt.

Hoofdstuk 3: Variant-gebaseerde uitspraakvariatie modellering voor automatische fonetische transcripties van spontane spraak

Het doel van de experimenten die zijn beschreven in dit hoofdstuk is het verbeteren van automatische fonetische transcripties van, specifiek, spontane spraak. Om dit doel te bereiken is getracht de uitspraakvariatie in spontane spraak beter te modelleren. In veel gevallen worden uitspraakvarianten van woorden door middel van regels gegenereerd en opgenomen in een lexicon. Een ASH systeem kiest vervolgens uit een dergelijk lexicon de variant die het beste past bij het spraaksignaal door middel van geforceerde herkenning. Op deze manier is het mogelijk een automatische fonetische transcriptie te maken. In spontane spraak echter, kan, zoals gezegd, de uitspraakvariatie nogal extreem zijn, wat het lastig maakt om die extreme verschillen binnen de variatie van één woord door middel van regels te modelleren. In dit hoofdstuk wordt een alternatieve modelleermethode getest waarbij uitspraakvarianten niet worden gegenereerd door middel van regels, maar door middel van extractie uit een groot, handgetranscribeerd corpus van spontane spraak (CGN). De meest frequente varianten met hun a priori waarschijnlijkheid worden geëxtraheerd en opgeslagen in het lexicon en een taalmodel. Het ASH systeem gebruikt vervolgens die informatie om de meest passende variant te kiezen. De automatische transcriptie die op deze variant-gebaseerde manier is verkregen, laat een verbetering zien van 6% (20% relatief) ten opzichte van een eenvoudige concatenatie van canonieke vormen. De verbetering mag aanzienlijk lijken, maar nog steeds komt de prestatie van de automatische transcriptieprocedure niet in de buurt van die van een menselijke transcribent.

Concluderend, de procedure die is beschreven en getest, is inderdaad geschikt om betere fonetische transcripties te genereren van spontane spraak. Tevens is aangetoond dat grote hoeveelheden spontane spraak met een handmatige transcriptie bijzonder nuttig zijn

om uitspraakvariatie in die spontane spraak te kunnen modelleren. Het is duidelijk dat, naarmate er meer data beschikbaar is, er beter gemodelleerd kan worden. Betere modellering leidt tot betere automatische fonetische transcripties, zodat uiteindelijk spontane spraak op een snellere en goedkopere manier getranscribeerd kan worden.

Hoofdstuk 4: Het meten van de kwaliteit van fonetische transcripties van grote spraakcorpora

In de meeste grote spraakcorpora zijn handmatige fonetische transcripties gegenereerd door middel van een procedure waarin een gegeven voorbeeldtranscriptie door een groep transcribenten wordt gecontroleerd en verbeterd. Op deze manier wordt niet alleen de benodigde tijd om transcripties te maken gereduceerd, maar ook de kosten. Meestal wordt de kwaliteit van de transcripties, die op deze manier gegenereerd zijn, gemeten door twee afzonderlijk gemaakte transcripties van een kleine sample met elkaar te vergelijken om zo de mate van overeenstemming tussen die twee transcripties te bepalen. In dit hoofdstuk wordt gepleit voor een additionele maat om transcriptiekwaliteit vast te stellen. Immers, als transcribenten een voorbeeldtranscriptie moeten verbeteren, is de kans aanwezig dat zij zich laten leiden door het gegeven voorbeeld en daarom vaker dan zou moeten de symbolen in het voorbeeld ongewijzigd laten. Het gevolg daarvan is een wellicht kunstmatig hoge overeenstemming tussen de transcribenten. Die hoge overeenstemming houdt niet automatisch in dat de symbolen in de transcripties ook daadwerkelijk het spraaksignaal representeren. Daarom wordt er voorgesteld om de kwaliteit vast te stellen door te vergelijken met een goede referentietranscriptie, in dit geval een consensustranscriptie, die de best mogelijke benadering is van de werkelijke uitspraak.

Er is een experiment opgezet waarin verschillende transcribenten een sample uit het CGN moesten transcriberen volgens bovenstaande procedure. Tegelijkertijd is een consensustranscriptie gemaakt, zonder een voorbeeldtranscriptie, door twee ervaren fonetici. Daarna zijn zowel de overeenkomsten tussen de afzonderlijke transcripties van de transcribenten, als de overeenkomsten tussen de afzonderlijke transcripties en de consensustranscriptie gemeten.

De resultaten laten zien dat de individuele transcribenten inderdaad de gegeven voorbeeldtranscriptie verbeteren, zodat de resulterende transcriptie de werkelijke uitspraak beter representeert. Maar tegelijkertijd komt ook naar voren dat de overeenkomsten tussen de transcribenten onderling groter zijn dan de overeenkomst tussen de afzonderlijke transcripties en de consensustranscriptie. Dit lijkt het vermoeden te bevestigen dat het gebruik van een voorbeeldtranscriptie de mate van overeenkomst tussen de transcribenten kunstmatig verhoogt. De werkelijke kwaliteit, in termen van nauwkeurigheid waarmee de uitspraak weergegeven wordt, zal daarom lager zijn. Daarnaast is gevonden dat de

toegevoegde waarde van de menselijke correctie van de voorbeeldtranscriptie niet bij alle spraakstijlen even groot is. Bij voorgelezen spraak geldt dat er wel een significante verbetering is gevonden ten opzichte van de voorbeeldtranscriptie zelf, maar dat er nog steeds systematische fouten worden gemaakt bij het beoordelen van de stemhebbende en stemloze fonemen. Om deze reden zouden toekomstige projecten, met het doel om transcripties te maken van grote hoeveelheden spraak, terughoudend moeten zijn met betrekking tot het inhuren van relatief dure menselijke transcribenten om voorgelezen spraak te transcriberen. Tot slot is aangetoond dat transcripties gemaakt volgens de procedure waarin een voorbeeldtranscriptie verbeterd moet worden, niet geëvalueerd kunnen worden door alleen de onderlinge overeenstemming te berekenen, maar dat een vergelijking met bijvoorbeeld een consensustranscriptie een meer valide beeld oplevert.

Hoofdstuk 5: Meerwoordsuitdrukkingen in gesproken taal

In hoofdstuk 3 is al gedemonstreerd dat grote hoeveelheden handgetranscribeerd spraakmateriaal zeer veel informatie bevatten die gebruikt kan worden om tot betere automatische fonetische transcripties te komen. In dit hoofdstuk wordt een groot corpus van spontane spraak gebruikt om uitspraakverschijnselen op een hoger niveau te analyseren, namelijk over woordgrenzen heen. De vraag die in dit hoofdstuk moet worden beantwoord, luidt: zijn er systematische patronen te ontdekken waarmee de extreme uitspraakvariatie, die voorkomt in spontane spraak, voorspeld kan worden? Als dit inderdaad het geval blijkt te zijn, dan kan die informatie gebruikt worden om uitspraakvariatie beter te kunnen modelleren ten behoeve van betere automatische fonetische transcripties.

Voor het experiment is om te beginnen een inventarisatie gemaakt van frequent voorkomende woordsequenties (N-grammen) op basis van de orthografische transcripties van spontane spraak uit het CGN. Ten tweede zijn voor een selectie van deze N-grammen de handmatige fonetische transcripties geanalyseerd. Uit die analyse bleek dat de uitspraak van de woorden in de N-grammen behoorlijk afwijkt van de canonieke representaties. Maar wat belangrijker is, is dat de uitspraak van de woorden in de N-gramcontext andere uitspraakpatronen laat zien dan wanneer dezelfde woorden zijn uitgesproken in andere contexten. Dit suggereert dat de N-grammen die geanalyseerd zijn, kunnen worden beschouwd als Meerwoordsuitdrukkingen en dat deze expressies in hun geheel zouden moeten worden opgenomen in een uitspraaklexicon met de daarbij behorende uitspraakvarianten. Op deze manier kan een ASH systeem tijdens geforceerde herkenning over informatie beschikken die over woordgrenzen heen gaat, waardoor verbeterde fonetische transcripties kunnen worden gegenereerd.

De resultaten behaald in dit onderzoek wijzen erop dat het gebruik van Meerwoordsuitdrukkingen een bijdrage kan leveren aan verbeterde fonetische transcripties van spontane spraak.

Hoofdstuk 6: Algemene discussie en conclusies

In het laatste hoofdstuk worden de conclusies van de experimenten uit de voorgaande hoofdstukken samengevat en bediscussieerd in het licht van gerelateerd onderzoek dat elders is uitgevoerd. Daarnaast worden suggesties gegeven om in toekomstig onderzoek de automatische fonetische transcripties te verbeteren van problematische spraakklanken, zoals het onderscheid tussen stemhebbende en stemloze fricatieven en het wel of niet aanwezig zijn van de schwa, /r/, /n/ en /t/ in het spraaksignaal.

Met betrekking tot het centrale thema van dit proefschrift – fonetische transcripties van grote spraakcorpora – is het duidelijk geworden dat fonetisch getranscribeerde spraakcorpora zeer waardevolle onderzoeksmiddelen zijn. Dankzij deze grote spraakcorpora is het mogelijk uitspraakverschijnselen te bestuderen en de opgedane kennis in te zetten ten behoeve van de generatie en evaluatie van verbeterde automatische fonetische transcripties. Kortom, onderzoek naar fonetische transcripties profiteert niet alleen van grote getranscribeerde spraakcorpora, maar draagt tegelijkertijd ook bij aan de ontwikkeling van nieuwe, nog uitgebreidere corpora die in de toekomst getranscribeerd moeten worden.

CURRICULUM VITAE

Diana Binnenpoorte was born on February 19th, 1974 in Nijmegen, the Netherlands. She spent her childhood in a small village near Nijmegen, named Hernen, where she enjoyed primary education at the Basisschool Heilig Hart. Subsequently, Diana attended the Pax Christi College in Druten from which she obtained her VWO diploma in 1993. In that same year she started as freshman at the University of Utrecht and studied General Arts. After she obtained her propaedeutic diploma, Diana started studying Linguistics and specialized in both Language and Speech Technology. In 1999, she received her Masters diploma after an internship at KPN Research in Leidschendam, which resulted in a thesis on speech technology in a mobile telephony environment. From March 1999 she worked as a junior researcher for the Speech Processing Expertise Centre, SPEX, and for the Department of Language and Speech, both at the Radboud University in Nijmegen. Diana worked on several (inter)national projects all concerning speech related research before she fully committed herself to the Dutch-Flemish CGN project (Corpus Gesproken Nederlands, *Spoken Dutch Corpus*) and left SPEX in November 2002. From March 2004, when the CGN project was finished, until August 2005, she dedicated her time on writing this PhD thesis which is based on parts of the work she carried out within the CGN project. At the moment Diana is working as a designer of spoken dialogue systems for the department Customer Contact Solutions at LogicaCMG in Nieuwegein.