VALIDATION, AUTOMATIC GENERATION AND USE OF BROAD PHONETIC TRANSCRIPTIONS

Cover design: Christophe Van Bael
Printed and bound by PrintPartners Ipskamp, Nijmegen

ISBN: 978-90-9022151-9

VALIDATION, AUTOMATIC GENERATION AND USE OF BROAD PHONETIC TRANSCRIPTIONS

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. mr. S.C.J.J. Kortmann,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op maandag 15 oktober 2007
om 13.30 uur precies

door
Christophe Patrick Jan Van Bael
geboren op 12 december 1978
te Mortsel, België.

Promotor	Prof. dr. L. Boves
Co-promotores	Dr. H. van den Heuvel Dr. H. Strik
Manuscriptcommissie	Prof. dr. R. van Hout (Voorzitter) Prof. dr. L. Pols (Universiteit van Amsterdam) Mevr. dr. M. Wester (Universiteit van Edinburgh)

Het onderzoek beschreven in dit proefschrift werd gesubsidieerd door de Stichting Spraaktechnologie, Utrecht, Nederland.

Acknowledgements

Before I came to Nijmegen, I had never pictured myself as a PhD student conducting academic research. Nonetheless, my time as a PhD student at the Department of Language and Speech of the Radboud University Nijmegen turned out much more enjoyable and rewarding than I had ever imagined. I owe this to a large extent to the following people.

Lou, I thank you for keeping up to date with research in so many fields, for sharing this knowledge with me, for supporting me throughout this project, and in particular also for the low threshold and the large opening hours to your room and your 24/7 connection to your mailbox.

Henk and Helmer, I thank you for your confidence and for giving me the academic freedom I have come to appreciate so much. The road towards this dissertation has not been the road we all had in mind when we set out on our journey. It has been an interesting road though, and a valuable experience.

Johan, Ambra, Annika and Andrea, thanks for being such interesting roommates. Special thanks go to Johan because without him explaining many neat programming tricks, my scripts would have run many times slower than they do now, and I would never have considered the department's coffee machine so easy to operate.

I also want to thank my other colleagues and ex-colleagues from the Department of Language and Speech for the many refreshing coffee breaks and lunches, and for the many birthday cakes and other treats. Thanks for all of the above, for making me feel at home from the start, and for limiting the number of jokes about Belgians to an average of one a day (this does not hold for you Folkert, I left you out of the equation in order not to skew the department's average too much).

Speaking of whom… Special thanks Folkert for your humour, the painful hours in the gym and the far less painful hours in the pub. Special thanks also go to Diana. Many thanks for your support and friendship and for helping me get a new job six months ahead of schedule. Great work! Folkert and Diana, I am very grateful that you agreed to become my 'paranimfen'.

As any other PhD student, I am indebted to the many people who helped me with my research and with preparing and improving my papers, presentations, and in the end also this thesis. Therefore, I want to thank again Lou, Henk and Helmer, as well as my other co-authors and helpful colleagues and ex-colleagues from the Department of Language and Speech, my co-authors from outside the department, the external reviewers of my papers, the manuscript committee and the attendants of my presentations at conferences and workshops.

Since I would not have been able to carry out this research and visit so many conferences and workshops without financial support, I sincerely thank the Stichting Spraaktechnologie and the International Speech Communication Association for their generous financial contributions to my project. Thanks for making this all possible.

Having thanked the academic people, I now turn to my family and friends who were always interested in and sometimes even intrigued (imagine!) by the things I was doing. Therefore, the rest of this section is in Dutch. Bedankt mama en papa, want jullie hebben mij de mogelijkheden gegeven en de discipline bijgebracht die nodig zijn om een meerjarenplan als dit te laten slagen. Bedankt Jo en Agnes, want de interesse en de geestdrift waarmee jullie me altijd gesteund hebben ervaar ik nog steeds als buitengewoon. Ik bedank graag ook mijn overige naaste (schoon)familie, in het bijzonder Patrick^{+}, en mijn vrienden in België. Het is fijn dat sommige dingen blijven zoals ze zijn. Ik krijg altijd weer een heerlijk gevoel als ik naar België rijd.

Tot slot bedank ik Anne. Anne, ik kan je niet genoeg bedanken voor je onvoorwaardelijke steun en liefde in de afgelopen vier en een half jaar, en in de vele jaren voor we naar Nijmegen kwamen. Jij bent zonder twijfel mijn beste vriendin, een fantastische vrouw en sinds kort ook een schitterende moeder.

Christophe Van Bael
Soest, 15 July 2007

Contents

Acknowledgements v

Contents vii

Chapter 1 – Introduction **1**

1.1 Phonetic transcriptions 2
1.1.1 A brief history 2
1.1.2 Different types of phonetic transcriptions 2
1.1.3 Usefulness of phonetic transcriptions for research and development 4

1.2 Generation of broad phonetic transcriptions 4
1.2.1 Manual generation of broad phonetic transcriptions 5
1.2.2 Automatic and semi-automatic generation of broad phonetic transcriptions 5

1.3 Validation of phonetic transcriptions 8
1.3.1 Terminology: validity and reliability 8
1.3.2 Traditional validation of phonetic transcriptions 10
1.3.3 An alternative to traditional validation procedures: Application-oriented validation 12

1.4 Aims and outline of this dissertation 13

Chapter 2 – Validation of Phonetic Transcriptions in the Context of Automatic Speech Recognition **15**

2.1 Introduction 16

2.2 Experimental setup 20
2.2.1 Traditional validation method 20
2.2.2 Application-oriented validation method 20

2.3 Material and continuous speech recogniser 23
2.3.1 Speech material 23
2.3.2 Phonetic transcriptions 24
2.3.3 Lexica 24
2.3.4 The continuous speech recogniser 25

2.4 Results and discussion 26
2.4.1 Traditional validation method 26
2.4.2 Application-oriented validation method 27

2.5 General discussion 29

Chapter 3 – Automatic Phonetic Transcription of Large Speech Corpora **33**

3.1 Introduction 34

3.2 Material and tools 36
3.2.1 Speech material 36
3.2.2 Canonical lexicon 37
3.2.3 Reference transcriptions 38
3.2.4 Continuous speech recogniser 38
3.2.5 Algorithm for dynamic alignment of phonetic transcriptions 39

3.3 Method 39
3.3.1 Generation of phonetic transcriptions with different transcription procedures 39
3.3.2 Evaluation of the phonetic transcriptions and the transcription procedures 46

3.4 Results 46

3.5 Discussion 49
3.5.1 Reflections on the evaluation procedure 49
3.5.2 On the suitability of a low-cost transcription procedure for the automatic phonetic transcription of large speech corpora 49
3.5.3 What about the remaining discrepancies? 53

3.6 Conclusions 54

Chapter 4 – Segment Deletion in Spontaneous Speech: A Corpus Study using Mixed Effects Models with Crossed Random Effects **57**

4.1 Introduction 58

4.2 Methodology 60
4.2.1 Data preparation 60
4.2.2 Analyses 62

4.3 Results 63
4.3.1 Frequencies of segment deletions 63
4.3.2 Modelling segment deletion 69

4.4 Conclusions 72

Chapter 5 – Speaker Classification by means of Orthographic and Broad Phonetic Transcriptions of Speech **75**

5.1 Introduction 76

5.2 Corpus material and transcriptions 77

5.3 Classification methodology 78
5.3.1 Classification algorithm 78
5.3.2 Classification variables 78

5.3.3 Classification features 80
5.3.4 Experimental setup 82
5.4 Classification results 83
5.4.1 Classification in terms of gender 83
5.4.2 Classification in terms of age 84
5.4.3 Classification in terms of regional background 85
5.4.4 Classification in terms of education level 86
5.4.5 More specific speaker classes 87
5.5 Conclusions and plans for future research 88

Chapter 6 – General Discussion **91**
6.1 Validation of broad phonetic transcriptions 92
6.2 Automatic generation of broad phonetic transcriptions 94
6.3 Research with broad phonetic transcriptions 96
6.3.1 Linguistic research with manually verified phonetic transcriptions 96
6.3.2 Linguistic research with automatic phonetic transcriptions 97

Bibliography 101

Summary 109

Samenvatting (Summary in Dutch) 115

List of Publications 123

Curriculum Vitae 125

CHAPTER 1

INTRODUCTION

1.1 Phonetic transcriptions

1.1.1 A brief history

Phonetic transcriptions are written representations of speech. Similar to the way *orthographic transcriptions* represent the spelling of words as strings of symbols called *graphemes*, phonetic transcriptions describe the pronunciation of words (i.e. the sequence of speech sounds or *phones* in words) as strings of symbols that are usually referred to as *phonemes* or *allophones*.

The study of phonetics and the use of phonetic transcriptions date back to around 1500 BC, when priests in different regions of India used phonetic transcriptions as an aid to preserve and propagate the original pronunciation of the Vedas, religious scriptures of the ancient Hindus (Kemp, 1994b). These early phonetic efforts were soon incorporated in the description of Sanskrit (Deshpande, 1994). Because of the diversity of phonetic transcription systems that emerged throughout history and because of the different aims for which these systems were developed and used (Ohala, 1994), it was not until the advent of the *Alphabet of the International Phonetic Association (IPA)* in 1888 (Kemp, 1994a) and computer readable alternatives such as the *ARPABET* and the *Speech Assessment Methods Phonetic Alphabet (SAMPA)* in the seventies and eighties of the previous century (Shoup, 1980; Wells, 1997) that phonetic transcriptions could be easily used in linguistic studies and speech applications that crossed language boundaries (Ohala, 1994). Figure 1.1 illustrates an orthographic transcription, an IPA transcription, an ARPABET and a SAMPA transcription of the sentence "At last, this dissertation is ready", spoken with a British accent.

orthographic transcription	at	last	this	dissertation	is	ready
IPA transcription	æt	lɑːz	ðɪz	dɪsəteɪʃn	əz	ɹɛdi
ARPABET transcription	aet	laaz	dhihz	dihsaxteyshn	axz	rehdiy
SAMPA transcription	{t	lA:z	DIz	dIs@teISn	@z	redi:

Figure 1.1: An orthographic, IPA, ARPABET and SAMPA transcription of the sentence: "At last, this dissertation is ready", spoken with a British accent.

1.1.2 Different types of phonetic transcriptions

Originally developed as a tool to preserve cultural and religious heritage, phonetic transcriptions gradually came to serve various other purposes. As a result, different transcription types emerged, and the term 'phonetic transcription' became a generic term covering transcriptions that can be described as:

- *systematic* or *impressionistic*, according to the purpose they are generated for
- *phonemic* or *allophonic*, according to the linguistic status of their phonetic symbols
- *broad* or *narrow*, according to the level of detail in their symbol set

(Ladefoged, 1993:40-42; Laver, 1995:549-562).

First explaining the difference between phonemic and allophonic transcriptions and subsequently the difference between broad and narrow transcriptions greatly facilitates the explanation of the systematic-impressionistic dichotomy.

Phonemic transcriptions represent speech in terms of the basic contrasting units or *phonemes* of a language's sound system. Phonemes can be defined as the speech sounds by means of which the speakers of a language can change the meaning of words. In English, for example, the words 'map' and 'nap' have a different meaning only because of the alternating use of the sounds /m/ and /n/. Therefore /m/ and /n/ are considered two different phonemes in English. Although large overlaps occur between the phoneme inventories of natural languages, each language is characterised by a specific inventory. In conversational speech, phonemes are often realised differently in different contexts. The phoneme /n/, for example, is mostly pronounced with the tip of the tongue touching the alveolar ridge behind the upper teeth. In most varieties of English, however, the same phoneme /n/ is pronounced with the tip of the tongue touching the upper teeth rather than the alveolar ridge when it occurs before a dental fricative such as in 'tenth'. The phoneme /n/ is often even pronounced as a bilabial instead of an alveolar or a dental nasal when it precedes a bilabial stop such as in 'brown bear'. Dental, bilabial and other realisations of the phoneme /n/ are called *allophones* of that phoneme when they can be attributed to the application of a structural phonological rule of the language to the underlying phoneme. Both the dental and the bilabial pronunciation of /n/, for example, are the result of a well-known English phonological rule called *regressive assimilation of place of articulation* which defines that a sound is articulated with a closure of the oral cavity at the same place as the closure made for the following sound. Transcriptions which represent the standard pronunciation of phonemes in words without reflecting any alternating pronunciation due to the application of phonological rules are called *phonemic transcriptions*. Transcriptions which represent the actual realisation of phonemes insofar they result from the application of phonological rules are called *allophonic transcriptions*. Phonemic transcriptions are useful to describe the systematic use of contrasting sounds in a language, but they cannot be used to represent the structurally defined pronunciation of words. For that purpose, allophonic transcriptions are used.

The number of different symbols with which transcriptions are generated determines the phonetic detail transcriptions can represent. Phonetic transcriptions that, because of the limited size of their symbol set, can only describe *broad* or general phonetic detail are called *broad (phonetic) transcriptions*. Phonetic transcriptions that are generated with a more elaborate symbol set, and that can therefore also describe *narrow* or fine phonetic detail are called *narrow (phonetic) transcriptions*. Since the phoneme inventories of all languages are limited in size, phonemic transcriptions can be considered broad phonetic transcriptions.

Allophonic transcriptions can be either broad or narrow transcriptions, depending on the number of different symbols they display.

Phonetic transcriptions are called *systematic transcriptions* if they represent the sound system or *phonology* of a language. Phonemic and allophonic transcriptions are systematic transcriptions because they represent only the contrastive sounds of a language or the phonologically determined variants of these sounds. As opposed to systematic transcriptions, *impressionistic transcriptions* are not phonologically but phonetically motivated. Their purpose is not to describe speech sounds in the phonological framework of a specific language, but rather to describe the pronunciation of phones with as much phonetic detail as their symbol set allows. *Impressionistic transcriptions* are often used to describe new languages or pathological speech for which no phonological assumptions can be made.

The diversity of transcriptions described here shows that "there is not ONE possible segmental [phonetic] transcription of a given utterance, but various ones, depending on the aim of the research" (Cucchiarini, 1993:3). For the purpose of this dissertation, it is important to note that it addresses research on and with systematic broad allophonic transcriptions, hereafter referred to as broad phonetic transcriptions of speech.

1.1.3 Usefulness of phonetic transcriptions for research and development

Phonetic transcriptions started out as a means of recording the pronunciation of words some 3400 years before the mechanical recording of sound was made possible through inventions of Scott (in 1857; his phonoautograph could not yet play back sound), Edison (in 1877; his phonograph could play back sound) and many others in the nineteenth and twentieth century (Straw, 1993). Phonetic transcriptions soon became part of the standard tool chest for descriptive linguistic research (Deshpande, 1994), and they have ever since proven useful in the fields of phonetics (Ladefoged, 2003), phonology (Labov, 1994; Ladefoged and Maddieson, 1996), sociolinguistics (Nerbonne et al., 1996), language pedagogy, lexicography (Wells, 2000) and the study of speech and language disorders and ensuing speech therapy (Howard and Heselwood, 2002).

In addition, the advent of computers in the seventies and the strong increase in computing power in the eighties and the nineties of the previous century created new computer-driven speech applications that required the availability of ever larger amounts of phonetic transcriptions. Nowadays, phonetic transcriptions are also used in computer assisted pronunciation training (Neri et al., 2002), in automatic speech recognition (Strik and Cucchiarini, 1999) and in text-to-speech synthesis (Bellegarda, 2005).

1.2 Generation of broad phonetic transcriptions

Broad phonetic transcriptions can be generated manually, automatically or semi-automatically. Whereas the manual production of broad phonetic transcriptions completely

relies on the efforts of human transcribers, automatic and semi-automatic transcription procedures require the use of an automatic transcription system, the output of which is either taken for granted (automatic transcription procedures) or manually verified and corrected by human transcribers (semi-automatic transcription procedures).

1.2.1 Manual generation of broad phonetic transcriptions

Obviously, manual transcription procedures have the longest history. Well before the advent of computers, people recorded speech by translating their auditory perception of the signal into sequences of phonetic symbols. Although human transcription quality is still generally considered a valuable reference for assessing the quality of automatic transcriptions (see Section 1.3.2), the manual generation of phonetic transcriptions has proven to be a time-consuming and therefore expensive undertaking. Binnenpoorte (2006) reported that the transcription of one minute of conversational speech takes about sixty minutes when the transcription is made from scratch. In addition, on various occasions, human transcribers have shown to make biased judgements because of their phonetic expectations of the speech signal (Oller and Eilers, 1975), their transcription training (Ladefoged, 1960; Catford, 1974) or the transcription protocol they try to adhere to. Even more disturbing because more difficult to monitor, human transcribers often make random errors, for example due to fatigue (Cucchiarini, 1993). This can all translate into intra-transcriber disagreements (the same stretch of speech is transcribed differently by the same transcriber at two points in time) and inter-transcriber disagreements (the same stretch of speech is transcribed differently by different transcribers) (Shriberg and Lof, 1991).

Whereas most previous phonetic, phonological and sociolinguistic studies were conducted on the basis of manually transcribed and therefore rather limited speech samples, computer-driven speech applications such as computer assisted pronunciation training, automatic speech recognition (ASR) and text-to-speech synthesis soon required much larger amounts of phonetically transcribed data. In the late eighties and the early nineties of the previous century, this increased need for phonetically transcribed speech data instigated the recording and phonetic transcription of large speech corpora such as the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Lamel et al., 1986; Fisher et al, 1986; TIMIT, 1990). Although TIMIT was still entirely transcribed by hand, soon a new line of research started investigating procedures to transcribe large speech corpora in an automatic or semi-automatic fashion (Binnenpoorte, 2006; Demuynck et al., 2004; Greenberg, 1997; Ljolje and Riley, 1991; Ljolje et al., 1997; Schiel, 1999; Vorstermans and Martens, 1994).

1.2.2 Automatic and semi-automatic generation of broad phonetic transcriptions

Both automatic and semi-automatic transcription procedures require the use of a computer-driven automatic transcription system, either to deliver an example transcription that can be

manually verified (semi-automatic transcription procedures) or to generate a transcription that can be considered as is (automatic transcription procedures). A*utomatic phonetic transcriptions* (APTs) can be generated in a number of ways.

Grapheme-based transcription procedures

One of the easiest ways to generate APTs is to substitute every word in the orthographic transcription with its *canonical transcription*, i.e. the transcription of the canonical or standard pronunciation of the word as if it is spoken in a formal style in isolation from the context of neighbouring words (Laver, 1995:551). The canonical pronunciation of words is usually stored in a pronunciation dictionary (or *lexicon*), an ordered list of words (represented as strings of graphemes) with their pronunciation (represented as strings of phonemes or allophones). The canonical pronunciation of words can be obtained manually through introspection or automatically through so-called *grapheme-to-phoneme* rules which substitute graphemes with phonemes (Bellegarda, 2005).

Because grapheme-based transcription procedures consider the graphemes in orthographic transcriptions to contain sufficient information to derive phonetic transcriptions from, the actual speech signal is ignored in the transcription process. This makes grapheme-based transcription procedures relatively easy to implement. However, the ignorance of the speech signal as a clue to determine the pronunciation of words and the exclusive use of one standard pronunciation per word imposes severe limitations on grapheme-based transcription procedures, the most important limitation being their inability to account for pronunciation variation.

This inability to transcribe the same word in different ways according to how it is actually pronounced is problematic because speakers tend to insert, delete and substitute speech sounds in real-life speech, particularly in more spontaneous speech (Cucchiarini and Binnenpoorte, 2002; Johnson, 2004). Transcription procedures that make their decisions on the basis of the acoustic properties of the speech signal instead of on the basis of the graphemes in the orthographic transcription alone, and that explicitly allow words to be transcribed in different ways do not pose such a problem. We call such procedures *signal-based transcription procedures.*

Signal-based transcription procedures

Signal-based transcription procedures usually require the use of an ASR system. Although ASR systems are usually built and employed for word recognition (i.e. automatic orthographic transcription) conventional ASR systems can be easily converted into systems that label smaller linguistic units such as phones. The transcription of phones can be done through *free phone recognition*, *constrained phone recognition* and *forced recognition* (in this context also referred to as *forced alignment)*. The names of these procedures refer to the degree of constraints on the search space of the ASR system. These constraints act as sign posts that guide the ASR system

to the most plausible symbolic representation of the acoustic signal. More guidance (or: the use of more constrained or restricted transcription methods) usually results in better transcriptions.

Free phone recognition is the least restrictive signal-based transcription method; it only requires an ASR system and a set of acoustic models representing the acoustic properties of the sounds in the language to be transcribed. Acoustic models of ASR systems are trained by providing such systems with labelled speech units. After being provided with enough correctly labelled units (words, phones or other units of speech), acoustic models represent the variation in the acoustic properties of those units. ASR systems running in free phone recognition mode do not employ linguistic constraints to limit their transcription options, hence the name of the procedure. ASR systems that perform free phone recognition only rely on their acoustic models to locate and label phones in speech.

Constrained phone recognition is a more restrictive transcription method because in addition to an ASR system with a set of trained acoustic models, it requires a phonotactic model that describes the (likelihood of) phone sequences in the language at hand. The use of phonotactic models puts a constraint on the search for the best phonetic symbol given the acoustic context, hence the name of the procedure.

Forced recognition is the most restrictive transcription method, because in addition to an ASR system with well-trained acoustic models, it requires an orthographic transcription of the material that has to be transcribed and a pronunciation lexicon with one or more pronunciations for every word in the orthographic transcription. ASR systems that perform forced recognition are literally *forced* to retrieve the best matching pronunciation variant in the pronunciation lexicon for every word in the orthography, hence the name of this procedure.

The generation of plausible pronunciation variants and the addition of such variants to pronunciation lexica are commonly known as *lexical pronunciation modelling*. As a subfield of ASR, lexical pronunciation modelling received much attention around the turn of the century. This is reflected in the organisation of various workshops on the topic (e.g. the 1998 ESCA Tutorial and Research Workshop on modelling pronunciation variation for automatic speech recognition in Kerkrade, the Netherlands, and the 2002 ISCA Tutorial and Research Workshop on Pronunciation Modelling and Lexicon Adaptation for Spoken Language in Colorado, USA), special issues of scientific journals (e.g. Strik, 1999) and the publication of various PhD dissertations on the topic (e.g. Kessens, 2002; Saraçlar, 2000; Wester, 2002).

Manual verification of automatic phonetic transcriptions

In practice, the automatic phonetic transcriptions of large speech corpora are sometimes at least partially verified and corrected (sometimes even double-checked) by human transcribers in an attempt to correct the most salient transcription errors. Examples of such corpora are the Buckeye Corpus of Conversational Speech (Pitt et al., 2005; Pitt et al., 2006) and the Spoken Dutch Corpus, a recent 9M word corpus of contemporary Dutch of which the automatic phonetic transcriptions of a 1M word subset were manually verified and corrected by trained linguistics students (Gillis et al., 2001; Goddijn and Binnenpoorte, 2003; Oostdijk, 2002).

The use of an automatically generated example transcription was found to increase the transcription speed considerably as opposed to manual transcription from scratch. Demuynck et al. (2002) reported that their students needed 40 minutes to verify the transcription of one minute of spontaneous speech (and 15 minutes to manually verify the transcription of one minute of public lectures). This is a considerable improvement to the aforementioned 60 minutes required to transcribe one minute of conversational speech from scratch (Binnenpoorte, 2006). However, semi-automatic transcription procedures also imply risks.

No matter how well linguists are trained, expert judgements of an example transcription can still be biased by the linguist's linguistic expectations (Coussé and Gillis (2006) found regional influences in the manually verified transcriptions of the Spoken Dutch Corpus), by the training the transcribers received, and by the protocol describing how and when to modify the example transcription. In addition, the use of an example transcription can tempt human transcribers into adhering to the example transcription despite contradicting acoustic cues in the speech signal. Demuynck et al. (2004), for example, reported cases where human transcribers preferred not to change the example transcription in the presence of contradicting acoustic cues, and cases where transcribers left phones in the example transcription that could not be aligned with a specific portion of the speech signal. Binnenpoorte (2006:9) therefore questioned "*the added value of having human transcribers correct an automatically generated phonetic transcription*".

Having described the benefits and the potential drawbacks of manual, automatic and semi-automatic transcription procedures, a description of a good procedure to evaluate phonetic transcriptions seems in place.

1.3 Validation of phonetic transcriptions

1.3.1 Terminology: validity and reliability

Although phonetics had already been practiced for centuries and although phonetic transcriptions had been used equally long, it was not until the second half of the previous century that linguists became aware of the importance of carefully evaluating their phonetic measurements before sharing their conclusions with others. Ladefoged's introduction to his 1960 paper on *The Value of Phonetic Statements* is quite illustrative in this respect:

> *"It is odd that linguists, who pride themselves on the rigor and scientific nature of many of their concepts, should nevertheless be so tolerant of vague, unverified statements in some parts of their field. To take an example, [Bernhard] Bloch has made many contributions to linguistic theory in a long series of excellent publications, but he does not appear to have adopted any scientific procedure to check the validity of the phonetic statements."*

Ladefoged, who also considered himself as one of "*the majority of linguists [who] constantly use loose unverified descriptions in all their work*" (Ladefoged, 1960:387) primarily referred to the more general notion of phonetic *statements* and not so much to phonetic *transcriptions* as such. Nevertheless, his message was clear: no interpretations of speech should be propagated without *validating* them first.

The term *validity* has always been important in statistics and the social sciences. In these fields, measures, instruments and experiments are considered valid when they measure what they are supposed to measure (Cucchiarini, 1993:11). Similarly, in the much younger field of speech corpora and annotation science, speech corpora and data annotations are considered valid when they meet the prescribed specifications in the corpus documentation (Van den Heuvel and Sanders, 2006). In a general way, the validity of a measure, an instrument, an experiment, a corpus or its annotations can be described as the extent to which they serve the purpose they are used for.

The term validity is closely linked to, but distinct from, the notion of *reliability*. The reliability of a measure, a test or an instrument refers to its consistency or its ability to yield the same results in repeated measurements of the same phenomenon under identical conditions. Whereas validity implies reliability (an experiment can only be valid if the same values are found for the same variables in repeated measurements under identical conditions), reliability does not imply validity (an experiment can be reliable in that it measures variables consistently that are nonetheless inconclusive or invalid to confirm or falsify a hypothesis).

In her study on the nature of variation in phonetic transcriptions, Cucchiarini (1993:12-13) described the terms validity and reliability in the context of phonetic transcriptions. She stated that "*The validity of a transcription could be estimated by comparing the transcription to the reality it is supposed to represent (the 'true' criterion scores).*" However, it was instantly added that "*this raises two questions: what is a transcription supposed to represent and how can 'true' criterion scores be obtained?*"

Phonetic transcriptions, as stated in the introduction to this Chapter, describe the pronunciation of words. But, since phonetic transcriptions are the written representations of perceptual analyses of speech, the accuracy with which they represent the speech signal is a priori restricted by the auditory perception of the transcriber, and by the detail of the symbol set with which the transcriptions will be generated. Each mapping of a continuous speech signal onto a sequence of discrete phonetic symbols taken from a finite symbol set implies some degree of quantisation error. These errors show in the time domain as well as in the acoustic domain, for all acoustic properties in a certain time interval have to be represented by just one symbol. Obviously, the quantisation errors in both domains will be larger if fewer symbols are used, as is the case with broad phonetic transcriptions. Considering this, phonetic transcriptions cannot represent all articulatory properties characterising the pronunciation of words, but only "*those articulatory properties of speech sounds which filter through auditory perception and the symbol system*" (Cucchiarini, 1993: 13). This means that phonetic transcriptions can be validated by comparing them to a reference transcription: a segmental, restricted description representing exactly *those articulatory properties*.

1.3.2 Traditional validation of phonetic transcriptions

As indicated in the previous Section, the validation of language resources (e.g. speech corpora) and their annotations is usually conducted by systematically comparing them with their specifications in the documentation of the resource (Van den Heuvel and Sanders, 2006). In the world of language resources such a 'validation procedure' serves as a quality assessment: language resources (and annotations) that are in agreement with their specifications are indirectly supposed to serve the purpose(s) they were generated for – and thus considered valid.

Nowadays speech corpora are often delivered with a lexicon comprising one or more pronunciation variants for every word in the orthographic transcription (e.g. the databases constructed in the SpeechDat framework - Van den Heuvel et al., 2001), or with a broad phonetic transcription which is provided as an additional annotation layer (e.g. Switchboard - Greenberg, 1997; The Spoken Dutch Corpus - Goddijn and Binnenpoorte, 2003). These transcriptions are usually validated by an independent human expert who assesses the acceptability of the symbols in a representative subset of the transcriptions in terms of their correspondence to the transcription guidelines coming with the language resource (TIMIT, 1990; Greenberg, 1997; Goddijn and Binnenpoorte, 2003; Pitt et al., 2005). The larger the number of agreements or the smaller the number of disagreements, the more valid phonetic transcriptions are considered to be. Figure 1.2 illustrates the alignment of the phonetic symbols in a manually verified phonetic transcription (PT) from the Spoken Dutch Corpus and a reference transcription (RT), and the subsequent computation of the number of symbol substitutions (sub), insertions (ins) and deletions (del) and the overall disagreement measure (percentage disagreement).

PT	E	n	d	A	d	b	@	v	i	l	w	-	-	a	r	d	@	x		
RT	E	n	d	a	-	b	@	v	i	l	w	E	l	a	r	d	I	x		
				sub	ins							del	del				sub			
Dutch	e	n	d	a	t	b	e	v	ie	l	w	e	l	aa	r	d	i	g		
English	and			that		was		pleasant						quite						

$$\%\ \text{disagreement} = \left(\frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{reference symbols}}\right) \times 100\% = \left(\frac{2+2+1}{17}\right) \times 100\% = 29{,}41\%$$

Figure 1.2: Alignment and computation of disagreement between a phonetic transcription (PT) and a reference transcription (RT).

Agreements and disagreements between a transcription and its reference transcription(s) can be counted by hand, but increased computing power has enabled us to automatically align

strings of phonetic symbols on the basis of their articulatory features, and to compute the number of identical and different symbols accordingly (Cucchiarini, 1993).

Since the auditory perception of human transcribers is open to subjectivity and since automatic phonetic transcriptions and semi-automatic transcriptions are not necessarily flawless either, it is generally considered impossible to generate an objective, 'true', ultimate reference transcription to compare and validate phonetic transcriptions with. The literature has provided useful approximations to the elusive 'true' transcription though, most of which were designed as aids for linguistic research rather than as reference transcriptions for validation purposes.

In their study on speech sound acquisition by three- to six-year-olds, Shriberg et al. (1984) describe the generation and the use of a *consensus transcription*. A consensus transcription is generated by two or more transcribers who are forced to agree on every symbol in the transcription. Whereas Shriberg et al. (1984) originally described the rules to generate and use a *narrow* consensus transcription, *broad* consensus transcriptions can be used for the validation of broad phonetic transcriptions (Goddijn and Binnenpoorte, 2003). The most important advantage of using a consensus transcription as a reference to compare and validate phonetic transcriptions with is the careful consideration of every speech sound by all transcribers, so that the resulting transcription will most probably be very plausible. The most important disadvantage is the time and money required to have several human transcribers agree on the symbolic representation of every speech sound. In addition to this disadvantage, the generation of consensus transcriptions involves the risk that one or a few transcribers dominate the decision making. This can introduce a bias in the transcription.

Majority vote transcriptions offer an alternative to consensus transcriptions as reference transcriptions for the validation of phonetic transcriptions. Kuijpers and van Donselaar (1997) generated such a majority vote transcription to study schwa insertion and schwa deletion in Dutch. Three trained transcribers transcribed the speech material individually. Their transcriptions were compared afterwards. In the case of discrepancies between the phonetic symbols in the transcriptions, the symbol used by two of three transcribers was retained. If there was no agreement between the transcribers, the transcription was excluded from the study. Although the authors used a majority vote procedure as a means of describing pronunciation variation, majority vote transcriptions could also be used as reference transcriptions to compare and validate other transcriptions with. Compared to the generation of consensus transcriptions, the generation of majority vote transcriptions as described in Kuijpers and van Donselaar (1997) cannot be dominated by one or more transcribers. In addition, the use of a high number of transcribers still gives a fair chance of a plausible reference transcription. However, the use of several transcribers (three or more in order to enable a majority vote when the transcribers disagree) also implies substantial time and monetary constraints, and it may not be easy for transcribers to reach a majority vote on every symbol if they have to make a complete transcription. Majority votes can be reached easier if transcribers have to attend to just one or a few speech processes in controlled contexts (e.g. the presence or absence of schwa in particular phonetic contexts).

As opposed to the aforementioned validation procedures which all involve the use of just one reference transcription, phonetic transcriptions can also be validated in terms of their (dis)agreements with several phonetic transcriptions made by different transcribers. Although such validation procedures are primarily used to assess the transcription consistency of and between individual transcribers, they also offer a multi-valued reference with which phonetic transcriptions can be compared: the disagreements between a transcription and each individual reference transcription can be compared with the pair wise disagreements between each of the reference transcriptions (Binnenpoorte et al., 2003; Greenberg et al., 1996; Kipp et al., 1996; Kipp et al, 1997). The validation of phonetic transcriptions in terms of various reference transcriptions has the same benefits and drawbacks as the use of a majority vote transcription: the employment of more transcribers may increase the robustness of the validation procedure against idiosyncratic errors in the reference transcriptions. Unfortunately, it also implies considerable time and budgetary investments.

It can be concluded that each of these validation procedures is time consuming and expensive because of the employment of human transcribers for the generation of a reference transcription. This explains why the phonetic transcriptions of large speech corpora are commonly validated through the assessment of just a small subset of the transcriptions which is considered to be representative for the rest of the corpus (TIMIT, 1990; Greenberg, 1997; Goddijn and Binnenpoorte, 2003; Pitt et al., 2005).

1.3.3 An alternative to traditional validation procedures: Application-oriented validation

Irrespective of their individual advantages and disadvantages, each traditional validation procedure in which phonetic transcriptions are validated through a comparison with one or more reference transcriptions ignores one important part of the definition of validity given in the first paragraph of this Section: *[…] the validity of a measure or an instrument determines to what extent the measure or the instrument serves the purpose it is used for.* Looking back at the aforementioned validation procedures, it appears that none of them considers the variety of purposes phonetic transcriptions can serve. Since phonetic transcriptions are mostly used as (and sometimes even specifically generated to be) instruments to serve a specific purpose or application (e.g. phonetic research, automatic speech recognition, text-to-speech synthesis or pronunciation training), it seems only natural to validate them in terms of the application they are meant to serve whenever that application is a given. After all, a phonetic transcription which resembles a reference transcription better may not always lead to a better performance of each application it will be employed in. Application-oriented validation procedures in which phonetic transcriptions would be validated in terms of the measure directly related to the application they have to serve (e.g. the Word Error Rate -the percentage of erroneously recognised words- for automatic speech recognition) would at least guarantee that the suitability of a transcription for that particular application is assessed in the best possible way.

1.4 Aims and outline of this dissertation

The research reported in this dissertation was conducted with three aims in mind. These aims are reflected in Chapters 2 to 5, which report studies that were conducted between October 2002 and March 2007 with colleagues at the Centre for Speech and Language Technology (CLST) at the Radboud University Nijmegen and at the Interfaculty Research Unit for Language and Speech, also located in Nijmegen.

Our first aim was to reconsider the general applicability of traditional validation procedures that validate phonetic transcriptions through a comparison with one or more human-made reference transcriptions. Experience shows that, in particular in the context of large speech corpora, phonetic transcriptions are usually validated through such procedures, irrespective of the procedure the transcriptions were generated with (manual, automatic or semi-automatic) and irrespective of the research and applications they will be used for. Since phonetic transcriptions are often used to train automatic speech recognition systems, and since the relationship between recognition performance and a transcription's resemblance to one or more reference transcriptions has never been proven, we conducted a study to test whether a traditional validation method offers a useful indication of a transcription's suitability for the training of automatic speech recognition systems. This study is reported in Chapter 2. We validated a basic canonical transcription and a manually verified phonetic transcription in terms of their resemblance to a consensus transcription and in terms of the application they would be used for: the training of an automatic speech recognition system. The relation between the outputs of the traditional and the application-oriented validation procedure was not straightforward. This has considerable implications for the future validation of phonetic transcriptions.

Our second aim was to investigate the applicability of fully automatic transcription procedures for the generation of human-like transcriptions of large speech databases. So-called *speech corpora* are often partially provided with a semi-automatic (viz. manually verified) phonetic transcription. Since the employment of human transcribers is time-consuming and expensive, we tested whether we could create human-like transcriptions by means of a fully automatic and therefore quicker and cheaper transcription procedure. Chapter 3 reports how we systematically compared the output transcriptions of ten automatic transcription procedures with a semi-automatic (manually verified) transcription of the Spoken Dutch Corpus. One automatic transcription resembled the manually verified transcription quite closely. This result is promising because the automatic procedure by means of which the transcription was generated now enables us to transcribe speech corpora that would otherwise be too large to be ever transcribed or verified by human transcribers.

Our third aim was to investigate the usefulness of both semi-automatic (manually verified) and automatic phonetic transcriptions as tools for the disclosure of linguistic knowledge in large speech corpora. Chapter 4 reports a study in which we used a semi-automatic transcription from the Spoken Dutch Corpus to investigate under what circumstances phones

and syllables are deleted in spontaneous Dutch. To this end, we linked information from the metadata and the orthographic and syntactic annotations of the Spoken Dutch Corpus to every phone and syllable in the semi-automatic phonetic transcription of the Spoken Dutch Corpus. Subsequently we fitted mixed-effect models with crossed random effects to automatically determine which linguistic and extra-linguistic factors favour the deletion of phones and syllables in spontaneous Dutch. Our results proved the value of mixed-effect modelling with crossed random effects for the simultaneous and automatic assessment of the degree to which various variables are related to a particular phenomenon, and they proved the semi-automatic transcriptions of the Spoken Dutch Corpus to contain a wealth of linguistic knowledge that can be easily accessed through the use of statistics.

Chapter 5 reports a study in which we investigated the use of automatic phonetic transcriptions for automatic speaker classification. We used a classification algorithm (Linguistic Profiling; van Halteren, 2004) to automatically learn characteristic speech habits from broad phonetic and orthographic transcriptions of speech from speakers whose gender, age, level of education and regional background were known. This knowledge was subsequently used to determine the gender, age, level of education and regional background of unknown speakers on the basis of the orthographic and broad phonetic transcription of their speech. In order to train the classification algorithm, we required more phonetic transcriptions than were manually verified in the Spoken Dutch Corpus. Therefore, we used the most optimal transcription procedure from Chapter 3 to automatically generate a phonetic transcription of all speech that fitted our experimental design. The results of our study are promising but as yet inconclusive, which has led us to formulate suggestions for targeted additional research.

Chapter 6 summarises the insights gained from the aforementioned studies on the validation, the automatic generation and the use of broad phonetic transcriptions. Special attention is paid to the implications of our studies for the validation and the automatic generation of phonetic transcriptions of large speech corpora.

CHAPTER 2

VALIDATION OF PHONETIC TRANSCRIPTIONS IN THE CONTEXT OF AUTOMATIC SPEECH RECOGNITION

Reformatted from:

Van Bael, C., Heuvel, H. van den, Strik, H. (in press). Validation of Phonetic Transcriptions in the Context of Automatic Speech Recognition. In: *Language Resources and Evaluation*.

Abstract

Some of the speech databases and large spoken language corpora that have been collected during the last fifteen years have been (at least partly) annotated with a broad phonetic transcription. Such phonetic transcriptions are often validated in terms of their resemblance to a handcrafted reference transcription. However, there are at least two methodological issues questioning this validation method. Firstly, no reference transcription can fully represent the phonetic truth. This calls into question the status of such a transcription as a single reference for the quality of other phonetic transcriptions. Secondly, phonetic transcriptions are often generated to serve various purposes, none of which are considered when the transcriptions are compared to a reference transcription that was not made with the same purpose in mind. Since phonetic transcriptions are often used for the development of automatic speech recognition (ASR) systems, and since the relationship between ASR performance and a transcription's resemblance to a reference transcription does not seem to be straightforward, we verified whether phonetic transcriptions that are to be used for ASR development can be justifiably validated in terms of their similarity to a purpose-independent reference transcription.

To this end, we validated canonical representations and manually verified broad phonetic transcriptions of read speech and spontaneous telephone dialogues in terms of their resemblance to a handcrafted reference transcription on the one hand, and in terms of their suitability for ASR development on the other hand. Whereas the manually verified phonetic transcriptions resembled the reference transcription much closer than the canonical representations, the use of both transcription types yielded similar recognition results. The difference between the outcomes of the two validation methods has two implications. First, ASR developers can save themselves the effort of collecting expensive reference transcriptions in order to validate phonetic transcriptions of speech databases or spoken language corpora. Second, phonetic transcriptions should preferably be validated in terms of the application they will serve because a higher resemblance to a purpose-independent reference transcription is no guarantee for a transcription to be better suited for ASR development.

Keywords: *Broad Phonetic Transcriptions, Validation, Automatic Speech Recognition.*

2.1. Introduction

Phonetic transcriptions are the written records of perceptual analyses of speech. They describe continuous speech signals as sequences of discrete phonetic symbols. These symbols can be chosen from small (more general) or large (more detailed) sets of symbols, depending on the

purpose the transcriptions are generated for. Transcriptions can be handmade, machine-made or they can be generated through a joint effort of man and machine.

Formally speaking, the validity of phonetic transcriptions indicates the adequacy with which the transcriptions represent the original speech signals, and as such also the adequacy with which the transcriptions serve the purpose which they will be employed for (Cucchiarini, 1993). However, the purpose for which transcriptions are made is not always unique nor always known in advance. Some of the speech databases and large spoken language corpora that have been collected during the last fifteen years (e.g. Switchboard (Godfrey et al., 1992; Greenberg, 1997) or the Spoken Dutch Corpus (Oostdijk, 2002; Goddijn and Binnenpoorte, 2003)) have been (at least partly) annotated with a phonetic transcription without knowing the specific purpose(s) the transcriptions would serve, since the corpora were explicitly aimed at serving a wide variety of research and development projects. In such contexts, phonetic transcriptions can only be validated by means of a purpose-independent validation criterion.

More often than not, phonetic transcriptions are validated through a comparison with some handmade reference transcription (RT) that is considered to be the most accurate representation of the speech signal that can be obtained with a given set of transcription symbols. In the literature several different instantiations of RTs have been used. Saraçlar et al. (2000) used a manual transcription that was independently produced by a phonetician. Kipp et al. (1996) used several independently produced manual transcriptions, each of which served as an independent reference. Kuijpers and van Donselaar (1997) also used several independently produced manual transcriptions, but they used them as a single reference by considering only the majority vote for every phonetic symbol. Shriberg et al (1984) argued that the best possible transcription is obtained by forcing two or more expert phoneticians to agree on each and every symbol in the transcription. A so-called 'consensus transcription' differs from a majority vote transcription in that the latter does not involve a negotiation phase during which individual transcribers may change their original transcript. Irrespective of the procedure through which a reference transcription is obtained, we will call the validation of phonetic transcriptions in terms of their resemblance to an RT the *traditional validation method*.

There are at least two methodological issues that raise questions about the traditional method for validating phonetic transcriptions. The first issue relates to the status of the RT as the 'true' representation of the original speech signal. Since speech signals are the result of continuous dynamic gestures of articulators, each mapping of such a continuous process onto a sequence of symbols that are taken from a finite symbol set implies some degree of quantisation error. These errors show in the time domain as well as in the acoustic domain because all acoustic properties in a certain time interval have to be represented by just one symbol. Obviously, the quantisation errors in both domains will be larger if fewer symbols are used. The decision on the number and the identity of the symbols is to some extent dependent on the phonetician's background. It can be concluded that there is no such thing as the "true" representation of a speech signal in the form of a sequence of discrete symbols (Cucchiarini, 1993). Consequently, the concept of a unique symbolic representation of a speech signal is elusive at best. The traditional validation method, however, always requires such a unique representation in the form of a reference transcription.

The second methodological issue is less obvious. It is related to the seemingly undisputable operationalisation of the concept of a transcription's validity in terms of the transcription's similarity with a purpose-independent reference transcription; there may not always be such a clear correlation between a transcription's similarity to a reference transcription and the transcription's suitability to serve a certain purpose. For example, no matter what the accuracy of a broad phonetic transcription may be, it will not be suitable for a phonetician who wants to represent the degree of diphthongisation of long vowels, simply because a broad phonetic transcription only reflects two extreme stages of diphthongisation: the process is either fully present or completely absent. For other applications, in which the detail in the phonetic transcription seems to correspond to the detail required by the application, the usefulness of the traditional validation method may be more difficult to estimate in advance. One such application is the development of automatic speech recognition (ASR) systems.

ASR development requires large speech databases or spoken language corpora with corresponding phonetic transcriptions for several different purposes, including the training of acoustic models and the construction of pronunciation lexica. It is intuitively reasonable to expect that acoustic models will be less polluted if they are trained on the basis of a 'better' transcription, and to think that words will be more accurately recognised if the recogniser's pronunciation lexicon comprises 'better' phonetic transcriptions. If we assume that transcriptions are 'better' if they are 'more similar' to a reference transcription, we assume that the traditional validation method is suitable for validating transcriptions that are to be used for ASR development.

Interestingly, however, the inverse relation between a transcription's resemblance to an RT and ASR performance does not hold. Kessens and Strik (2004) investigated the relationship between the performance of a set of continuous speech recognisers, and the resemblance between an RT and phonetic transcriptions that were generated by the different recognisers. They concluded that recognisers with a higher recognition performance (or: a lower word error rate (WER)) do not guarantee the generation of phonetic transcriptions that are more similar to a given RT.

Since the relationship between recognition performance and a transcription's resemblance to an RT does not seem to be straightforward, this study was aimed at testing whether the traditional validation method offers a useful indication of a transcription's suitability for basic ASR development. If, in addition to the results in Kessens and Strik (2004), we would fail to find a positive relationship between a transcription's resemblance to an RT and its suitability to develop ASR systems, this would indicate that phonetic transcriptions may be better validated through an *application-oriented validation* method (which, in our case, would mean in terms of their contribution to ASR performance). Such a result would also indicate that ASR developers could save themselves the tedious and expensive effort of collecting reference transcriptions in order to validate phonetic transcriptions that may come with a new training database.

We required two resources to assess the validity of phonetic transcriptions in terms of their contribution to ASR performance. First, we required a corpus suitable for the training

and the evaluation of an ASR system. This corpus had to contain at least two different transcriptions that could be used for that purpose. Second, we needed a fixed platform to develop and test the ASR system, in order to isolate the effect of the phonetic transcriptions from the multitude of other factors that could affect the performance of the ASR system.

Our first requirement was met by the Spoken Dutch Corpus (Oostdijk, 2002), a 9-million-word spoken language corpus, 10% of which comes with a manually verified broad phonetic transcription (Goddijn and Binnenpoorte, 2003). The second type of transcription that we used, viz. a canonical representation, is available in the canonical lexicon that typically comes with every corpus for ASR development. The corpus and the two transcriptions are described in more detail in Sections 2.3.1 and 2.3.2.

The requirement of a fixed platform to isolate the transcriptions as the only factor affecting the recognition performance was met by fixing the training and test corpora as well as the language models of our system. As a consequence, we could study the effect of the two transcription types in relation to 1) the amount of phonetically transcribed material that was used to train the acoustic models (since the production of manually verified transcriptions is time-consuming and expensive, the amount of training speech that comes with a manually verified phonetic transcription cannot be expected to be as large as the amount of speech that can be annotated with a canonical representation), 2) the procedures with which the acoustic models were trained (with the canonical representations, the manually verified phonetic transcriptions, or through a bootstrap procedure involving both transcription types), and 3) the pronunciations in the recognition lexicon (canonical representations or manually verified phonetic transcriptions).

Since we aimed at investigating the direct influence of the two transcriptions in a fixed experimental design, we did not aim at optimising recognition performance by all possible means. Rather, our intention behind the fixed experimental design was similar to the intention behind the research conducted in the framework of the AURORA project, where the ASR decoder was fixed, and performance improvements could only be obtained by adapting the acoustic features (Pearce, 2001). For the same reason, it should be clear that we did not aim at generating the most *accurate* transcription possible. Rather, we aimed at testing whether the traditional and the application-oriented validation method agreed on their assessments of the validity of the phonetic transcriptions in order to establish whether the traditional validation method guarantees an adequate indication of a transcription's suitability for ASR development.

This paper is organised as follows. Section 2.2 describes how canonical representations and manually verified phonetic transcriptions were validated in terms of the traditional validation method and in terms of their contribution to recognition performance. Section 2.3 presents the speech material and the architecture of the speech recogniser. In Section 2.4, we present and discuss the results of the validation experiments. In Section 2.5, we discuss the implications of our results.

2.2 Experimental setup

We validated canonical representations and manually verified phonetic transcriptions (MPTs) of data comprising two different speech styles: read speech and telephone dialogues. The details of the transcriptions are presented in Section 2.3.2. Here we confine ourselves to mentioning that the canonical representations were generated by concatenating the standard pronunciations of the words in the orthographic transcriptions. The MPTs were made by trained students who checked and corrected canonical representations by listening to the speech signal. The reference transcriptions were consensus transcriptions produced by two trained phoneticians.

2.2.1 Traditional validation method

We compared the canonical representations and the manually verified phonetic transcriptions with reference transcriptions of the same data. To that end we aligned the transcriptions of every speech style with the appropriate RT. Subsequently we summarised the disagreements between the transcriptions and the RT in an overall disagreement measure that was defined as:

$$Percentage\ \ disagreement = \left(\frac{Sub_{phone} + Del_{phone} + Ins_{phone}}{N_{phone}} \right) \times 100\% \qquad (1)$$

i.e. the sum of all phone substitutions (Sub_{phone}), deletions (Del_{phone}) and insertions (Ins_{phone}) divided by the total number of phones in the RT (N_{phone}).

We used Align (Cucchiarini, 1996) to align the phonetic transcriptions and to compute the percentage disagreement between them. Align is a dynamic programming algorithm designed to compute the optimal alignment between two strings of phonetic symbols according to matrices in which the articulatory feature values for the phonetic symbols are defined. The optimal feature matrices were determined in previous research on similar data (Binnenpoorte and Cucchiarini, 2003). The matrices are presented in Appendix 2.1.

2.2.2 Application-oriented validation method

We validated the canonical representations and the MPTs in terms of their contribution to the overall recognition performance of a standard continuous speech recogniser. We adhered to the traditional evaluation metric for recognition performance in ASR, the word error rate (WER), which is defined as:

$$WER = \left(\frac{Sub_{word} + Del_{word} + Ins_{word}}{N_{word}} \right) \times 100\% \quad (2)$$

i.e. the sum of all word substitutions (Sub_{word}), deletions (Del_{word}) and insertions (Ins_{word}) divided by the total number of words in the orthographic reference transcription (N_{word}).

The overall recognition performance of a continuous speech recogniser can be influenced by numerous factors. Two important factors, viz. the quality of the acoustic models and the degree to which the pronunciation lexicon contains realistic phonetic transcriptions for words to be recognised, are directly dependent on the availability of suitable phonetic transcriptions. The quality of acoustic models depends on the suitability of the phonetic transcriptions of the training material, because acoustic model training involves a time-alignment of large amounts of speech with corresponding phonetic transcriptions. Likewise, the quality of a pronunciation lexicon is determined by the quality of its transcriptions, in that more realistic phonetic transcriptions increase the chance of words to be correctly recognised. In addition, it has repeatedly been found that recognition performance also depends on the (lack of) correspondence between the transcriptions in the recognition lexicon and the transcriptions with which the acoustic models are trained. As already indicated, we validated the canonical representations and the MPTs in terms of overall recognition performance. By fixing the continuous speech recogniser but for the acoustic models and the recognition lexicon, we guaranteed that differences in the overall recognition performance could only result from the transcriptions' influence on the acoustic models and the recognition lexicon.

Per speech style, we conducted a series of four experiments. In these experiments, we trained the same recogniser with different sets of acoustic models (all context-independent models with a fixed model topology, but trained with different transcriptions and different amounts of training data) and we tested the recogniser with different recognition lexica. Table 2.1 presents a schematic overview of the four experiments. The experiments were characterised by three variables: 1) the amount of training data we used to train the acoustic models (large or small training set), 2) the (combinations of) transcriptions we trained the acoustic models with (canonical, MPT or a bootstrap procedure involving both transcription types – see below) and 3) the type of the transcriptions in the recognition lexica (canonical or MPT).

Table 2.1: Overview of the recognition experiments.

	size of the training sets	transcriptions for the training of acoustic models	transcriptions in the recognition lexica
experiment 1	small	canonical	canonical
experiment 2	small	MPT	MPT-based
experiment 3	large	canonical	canonical
experiment 4a	large	bootstrap MPT + canonical	canonical
experiment 4b			MPT-based

In experiment 1, we trained acoustic models with the canonical representations of the small training sets (see Section 2.3.1), and we used the same transcriptions to build canonical recognition lexica. The results of the first experiment formed a good baseline for the second experiment, in which we used the MPTs of the same small training sets to train the acoustic models and to build MPT-based recognition lexica. Since the production of MPTs tends to be time-consuming and expensive, larger sets of MPTs than the ones used in this second experiment are hardly ever available.

The third experiment resembled the first experiment, in that we trained acoustic models with canonical representations and in that we used the same canonical recognition lexica. However, this time we trained acoustic models with the canonical representations of much larger amounts of training data. The increased size of the data sets (as opposed to the first experiment) had to provide insight into the importance of the size of data sets for the training of efficient acoustic models. All acoustic models used in the first three experiments were generated from scratch (i.e. starting from a linear segmentation of the material).

In ASR, one often uses modest amounts of MPTs to train initial sets of acoustic models that, in a second training pass, are further trained with larger amounts of automatic phonetic transcriptions. This training method is called bootstrapping. We applied bootstrapping since we assumed that acoustic models that were initially trained with a small amount of MPTs and that were subsequently further trained with a large amount of canonical representations would outperform acoustic models that were trained from scratch with only canonical representations.

In the fourth experiment, we used the acoustic models of experiment 2 (which were trained on the MPTs of the small data sets) to align the speech data of the large data sets with the corresponding canonical representations of the data. Then we trained new acoustic models with the time-aligned canonical representations of the large data sets. Since the resulting acoustic models were based on a two-pass training procedure with MPTs and canonical representations, recognition experiments were carried out with both the canonical recognition lexica (exp. 4a) and the MPT-based lexica (exp. 4b). The alternating use of these recognition lexica (while using the same acoustic models) enabled us to study the effect of the different types of transcriptions in the recognition lexica in isolation.

To conclude, these experiments allowed us to validate the canonical representations and the manually verified phonetic transcriptions in terms of their suitability to train acoustic models and to generate recognition lexica. The transcriptions' suitability was reflected in and measured in terms of the recogniser's overall recognition performance. Whereas experiments 1 and 2 provided insight into the general influence of the two transcription types on the recognition performance, experiments 1 and 3 assessed the influence of different amounts of training data on the training of efficient acoustic models. Experiments 4a and 4b allowed us to investigate the influence of the different recognition lexica on the recognition performance.

2.3 Material and continuous speech recogniser

2.3.1 Speech material

We extracted the speech material for our experiments from the Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN, 2004; Oostdijk, 2002). The Spoken Dutch Corpus is a 9-million-word multi-purpose spoken language corpus comprising Dutch as spoken in the Netherlands and Flanders in different communicative settings. The whole corpus was orthographically transcribed, lemmatised, and supplied with part-of-speech tagging. A 1-million-word subset of the corpus, the so-called *core corpus*, was enriched with a manually verified broad phonetic transcription and a syntactic annotation.

We conducted our experiments on speech from the Netherlands. The data we used comprised two speech styles with different acoustic and communicative properties: read speech (read aloud texts from a library for the blind) and conversational telephone dialogues. The read speech was recorded with table-mounted microphones and sampled at 16 kHz with a 16-bit resolution. The material comprised monologues with a vivid prosodic structure (due to the material's fictional content and the purpose the texts were read for: entertainment). The telephone dialogues were recorded through a telephone platform and sampled at 8 kHz with an 8-bit A-law coding. The two speakers in each conversation were recorded on separate channels.

Table 2.2: Statistics (number of words/tokens) of the data sets.

speech style		reference sets	experimental sets			
			large training set	small training set	development test set	evaluation test set
read speech	# words	1,108	532,451	47,517	7,940	7,940
	hh:mm:ss	0:04:57	44:55:59	4:04:28	0:40:10	0:41:39
telephone dialogues	# words	363	263,501	41,736	6,953	6,955
	hh:mm:ss	0:01:26	18:20:05	1:29:23	0:30:02	0:29:50

Per speech style, we divided the material into two separate data sets which will hereafter be called the *reference sets* and the *experimental sets* (see Table 2.2). The data in the reference sets were provided with a consensus transcription. This enabled us to validate the phonetic transcriptions according to the traditional validation method. The data in the experimental sets were used to validate the phonetic transcriptions in terms of their suitability for ASR development (a more application-oriented validation method). To this end, the transcriptions were used to train (large and small training sets), tune (development test sets) and test (evaluation test sets) our continuous speech recogniser. Except for the training sets (the large training sets comprised the small training sets), all data sets were mutually exclusive.

2.3.2 Phonetic transcriptions

We worked with broad phonetic transcriptions of speech. All transcriptions were generated with the CGN phone set comprising 46 phones. However, not all of these phones occurred frequently enough in the training data to train robust acoustic models. In order to alleviate this problem, we mapped the phones in the transcriptions to the 39 phones presented in Appendix 2.2.

The canonical representations were generated by means of a lexicon-lookup procedure in which every word in the orthography was substituted with its standard pronunciation as represented in the canonical pronunciation lexica described in Section 2.3.3.

We extracted the MPTs of the data in the reference sets, the small training sets and the development and evaluation test sets from the CGN. The MPTs of the CGN are based on canonical representations to which all obligatory word-internal phonological processes (such as assimilation and degemination) were applied (Goddijn and Binnenpoorte et al., 2003; Booij, 1999). Cross-word processes were not applied. Human transcribers verified and corrected these example transcriptions according to a strict protocol. They were instructed to change the automatic transcriptions only if they were certain that the changes would yield a transcription that was substantially closer to the actual speech signal. As a consequence, the MPTs of the CGN may have a bias towards the canonical representations. However, such a check-and-correct procedure is a standard transcription procedure that has also been followed in other transcription projects (e.g. Greenberg, 1997).

The RTs were made in a fundamentally different way. Whereas the MPTs were made by human transcribers manually verifying an automatically generated transcription, the RTs were generated by two expert phoneticians transcribing from scratch. The transcribers had to reach a consensus on every symbol in the RTs. As a consequence, our reference sets were quite small compared to the evaluation test sets. However, whereas consensus transcriptions are always limited in size, they are often used to assess the validity of transcriptions obtained by means of other transcription procedures (like the MPTs and the canonical representations in our experiments).

2.3.3 Lexica

Canonical pronunciation lexica

Our canonical lexica (one for each speech style) comprised one canonical pronunciation for every word in the development, evaluation and small training sets. The canonical lexica were compiled from the TST-lexicon (in-house version of 29-09-2004) and the CGN-lexicon. The TST-lexicon is a comprehensive multi-purpose lexicon for language and speech processing. It was compiled by merging various existing electronic lexical resources such as CELEX (Baayen et al, 1995; CELEX Lexical Database, 2005), RBN (Referentiebestand Nederlands, 2005), and PAROLE (PAROLE lexicon, 2005). The CGN lexicon (delivered with the first

release of the CGN) comprised the canonical representations of almost all unique word forms occurring in our data sets. The phonetic representations in the CGN lexicon were generated by means of TREETALK (Hoste et al., 2000), a grapheme-to-phoneme converter trained on the CELEX Dutch database (Baayen et al., 1995). Obvious errors in frequent words were manually corrected. The transcriptions of English loan words that were not yet included in the CGN lexicon were obtained from the CELEX English database (Baayen et al., 1995). The missing transcriptions of geographical names were obtained from ONOMASTICA (Quazza and van den Heuvel, 2000). The remaining out-of-vocabulary words were transcribed by means of a rule-based grapheme-to-phoneme converter (Kerkhoff and Rietveld, 1994) and the transcriptions were manually verified.

Pronunciation lexica with manually verified phonetic transcriptions

The MPT-based lexica (one for each speech style) were generated through word-to-transcription mappings between the orthographic transcriptions and the MPTs of the data in the development, evaluation and small training sets. We included the manually verified pronunciations of the words in the development and evaluation sets because not all of these words occurred in the small training sets. In doing so, we excluded the number of out of vocabulary words as an extra variable from the comparison of the canonical and the MPT-based lexica. Similarly, in order to exclude the lexical confusability from the comparison of the lexica, we retained only the most frequently observed pronunciation variant per word. This way both the canonical and the MPT-based lexica contained precisely one pronunciation for every word in the orthographic transcriptions.

The major difference between the canonical lexica and the MPT-based lexica was that the canonical lexica reflected the underlying morphological structure of the words and hypotheses about their underlying phonemic representations, whereas the MPT-based lexica mainly reflected knowledge about the most frequent pronunciation of the words in everyday speech. The MPT-based and the canonical lexica for the read speech contained different transcriptions for 40% of their entries, the lexica of the telephone dialogues for 45% of their entries.

2.3.4 The continuous speech recogniser

The continuous speech recogniser was built with the HTK toolkit (Young et al., 2001) using standard procedures. The characteristics of the recogniser were fixed in all experiments, except for the recognition lexicon and the acoustic models, which were based on the different phonetic transcriptions under investigation.

Several pre-processing procedures were applied to the speech signal. First pre-emphasis was applied. Feature extraction was implemented as a Fast Fourier Transform using a Hamming window every 10 ms for 25-ms frames. The mel-scaled filter bank analysis (50-8000 Hz for the read speech and 80-4000 Hz for the telephone dialogues) resulted in 39

cepstral coefficients per frame (12 coefficients and a separate energy component, and their delta and acceleration coefficients).

The recogniser used one back-off bigram language model per speech style. The evaluation test set perplexity of the read speech was 61.12. The evaluation test set perplexity of the telephone dialogues made 43.22. The lower test set perplexity of the telephone dialogues reflects the high frequency of standard phrases in the conversations. The higher test set perplexity of the read speech reflects the fact that the read speech comprised fragments with varied content from a number of different novels that were written by different authors. The order of magnitude of the test set perplexities was low enough to obtain credible WERs and at the same time high enough to not obscure the effects of improved acoustic models.

The acoustic models were 3-state continuous density left-right context-independent Hidden Markov Models. We trained speech style specific acoustic models on the canonical representations and the MPTs of the large and small training sets. Per set, 39 models were trained: 37 phone models, one model representing long silences, and one 1-state model modelling the optional short pauses between words (see Appendix 2.2). All models were gender-independent and accent-independent and comprised 32 mixture components (diagonal variance vectors) per state.

2.4 Results and discussion

2.4.1 Traditional validation method

Table 2.3 reflects the validity of the phonetic transcriptions of both speech styles as assessed in terms of their overall disagreement (in % disagreement) with a reference transcription.

Table 2.3: Validation of phonetic transcriptions in terms of their deviation from a reference transcription. The lower the % disagreement, the better the transcription is considered to be.

speech style	PT	substitutions (%)	deletions (%)	insertions (%)	% disagreement
read speech	canonical	7.39	3.51	1.14	12.04
	MPT	3.88	1.19	0.69	5.76
telephone dialogues	canonical	9.60	10.92	1.08	21.61
	MPT	4.68	2.64	1.08	8.4

The results in Table 2.3 are very clear: 1) the MPTs consistently resembled the RTs more than the canonical representations did ($p < .01$, t-test), and 2) the deviations of the different transcriptions from the RTs were larger when more spontaneous speech was involved. The significance of the differences suggests that the power of the test was sufficiently large despite the moderate size of the reference sets.

The relatively high resemblance between the MPTs and the RTs (as compared to the resemblance between the canonical representations and the RTs) is probably due to the fact that the MPTs and the RTs, even though produced according to different protocols (cf. Section 2.3.2), were produced by human transcribers who based their judgments on the actual speech signal. The canonical representations were automatically produced without taking the actual speech signal into account.

The results in Table 2.3 are in line with results published in the field. Binnenpoorte et al. (2003) also reported that the degree of resemblance between phonetic transcriptions and a reference transcription is inversely related to the degree of spontaneity of the transcribed speech, and proportional to the amount of manual effort devoted to the production of the transcriptions.

In any case, the results in Table 2.3 indicate that according to the traditional validation method, the validity of the MPTs of the Spoken Dutch Corpus is significantly higher than the validity of the canonical representations of the same material.

2.4.2 Application-oriented validation method

Table 2.4 reflects the validity of the phonetic transcriptions of both speech styles as assessed in terms of the transcriptions' contribution to recognition performance (in WER).

Table 2.4: Validation of phonetic transcriptions in terms of their influence on recognition performance. The lower the WER, the more suitable the transcription is considered to be.

	speech style	substitutions (%)	deletions (%)	insertions (%)	WER (%)
experiment 1	read speech	7.68	2.85	0.82	11.35
	tel dialogues	33.43	17.12	2.60	53.16
experiment 2	read speech	7.95	2.07	1.27	11.28
	tel dialogues	33.56	16.97	2.56	53.09
experiment 3	read speech	7.61	2.17	0.96	10.73
	tel dialogues	32.47	17.97	2.13	52.57
experiment 4a	read speech	7.36	2.75	0.91	11.01
	tel dialogues	33.64	16.99	2.66	53.30
experiment 4b	read speech	7.77	2.07	1.12	10.96
	tel dialogues	33.26	17.11	2.52	52.42

The modest nature of the recognition results in Table 2.4 can be partly explained by the lively prosody and fictional content characterising the read speech, and by the spontaneity and acoustic conditions characterising the telephone dialogues. Moreover, only bigram language models and context-independent acoustic models were used, since our main target, viz. validating phonetic transcriptions for ASR, only required the development of a standard recogniser that differed with respect to 1) the amount of phonetically transcribed data used to

train the acoustic models, 2) the type of transcriptions of the training data, and 3) the type of transcriptions in the recognition lexicon. It is most striking that for both speech styles, none of the experiments yielded significantly different WERs ($p > .05$, *t*-test).

The recognition results of the first two experiments imply that the canonical representations were as suitable as the MPTs for training acoustic models on relatively small data sets (40K words), and for building pronunciation lexica for recognition. Remarkably, this did not only hold for the read speech, but also for the more spontaneous telephone dialogues in which the actual pronunciation could be expected to differ substantially from the canonical representation of the words. The MPT-based ASR system obtained a WER of 53.09%, which was almost identical to the 53.16% WER obtained by the system that was developed on the basis of the canonical representation of the words.

A comparison of the results of the first and the third experiment illustrates that the use of larger training sets (500K) decreased the WERs, though not significantly (0.62% absolute decrease on the read speech, 0.59% absolute decrease on the telephone dialogues). We did not conduct a similar experiment with MPTs, since the Spoken Dutch Corpus does not provide MPTs for such a large training set (nor does any other corpus available to date). However, MPTs of smaller data sets can be used to train acoustic models which in turn can be used to get good initial segmentations of much larger data sets. In our fourth experiment, we validated MPTs and canonical representations in terms of their potential for such a bootstrapping procedure.

In experiment 4a, we used the acoustic models trained on the MPTs of the small data sets (experiment 2) to get good initial segmentations of the large data sets. These segmentations were generated through a forced alignment of the canonical representations with the speech signal. A comparison of the results of experiments 3 and 4a illustrates that the bootstrapping procedure did not yield significantly different recognition results.

A comparison of the results of experiments 4a and 4b shows that the combined use of the MPT-based lexicon and the bootstrapped acoustic models yielded better (though not significantly better) results than the use of the canonical recognition lexicon with the same models. Especially the recognition of the telephone dialogues was facilitated by the use of the MPT-based lexicon. This is probably due to a larger mismatch between the actual data and the canonical representation of the spontaneous telephone speech.

At last, a comparison of the results of experiments 1 and 2 on the one hand and experiments 3, 4a and 4b on the other hand indicates that for both speech styles the acoustic models trained on the small data sets could not be improved substantially by adding more training material.

Overall, our recognition results are in line with a similar study on spontaneous telephone dialogues in American English (Switchboard) by Saraçlar et al. (2000). In that study, recognition experiments were conducted with different sets of acoustic models (trained on MPTs and automatic phonetic transcriptions) and matching decision tree-based pronunciation models. Their results showed that acoustic models trained on human transcriptions (Greenberg, 1997) did not give lower WERs than acoustic models trained on canonical base forms. Saraçlar et al. (2000) found that the models trained on the MPTs gave lower phone

error rates, but no lower WERs than the models trained on the canonical base forms. They concluded that their results must have been due to the increased lexical confusability in the corresponding MPT-based recognition lexicon. Our results suggest that this cannot be the full explanation. By allowing only the most frequent transcription per word, we minimised the risk of increasing the lexical confusability. Still we observed similarly remarkable recognition results which seem to suggest that for our ASR task, the canonical representations served their purpose as well as the manually verified phonetic transcriptions.

2.5 General discussion

This study was aimed at investigating whether the validity (or: the suitability) of phonetic transcriptions for basic ASR development can be assessed by means of the traditional validation method, i.e. in terms of the transcriptions' deviations from a handmade reference transcription. Previous research (Kessens and Strik, 2004) has shown that the relationship between recognition performance and a transcription's resemblance to an RT should not be taken for granted. In order to evaluate the usefulness of the traditional validation method, we conducted a series of experiments in which we assessed the influence of two different types of transcriptions (canonical representations and manually verified phonetic transcriptions) of two different speech styles (read speech and telephone dialogues) on the overall recognition accuracy of a continuous speech recogniser. As opposed to the traditional validation method, the assessment of the transcriptions' suitability for one particular purpose can be considered as an application-oriented validation method.

The outcome of the traditional validation method (which did not take into account the purpose the transcriptions would be used for) was quite outspoken: the validity of the MPTs was assessed much higher than the validity of the canonical representations because the MPTs deviated much less from the reference transcriptions than the canonical representations did. The application-oriented validation method gave quite another estimate of the transcriptions' validity. The assessment of the transcriptions' suitability for ASR showed that the use of MPTs and canonical representations did not yield significantly different recognition performance. This implies that both the MPTs and the canonical representations were equally valid for the purpose of developing an ASR system.

A comparison of the outcomes of the two validation methods supports different conclusions. First of all, it should be stressed that the application-oriented validation method did not contradict the usefulness of MPTs for ASR development, since we did not get better recognition results when using the canonical representations for this purpose. Logically, this also implies that the application-oriented validation method did not contradict the usefulness of manually verified transcriptions as such. As a matter of fact, for other purposes than training straightforward ASR systems (e.g. training more elaborate ASR systems), the story may well be different. For applications such as research in phonetics, it will probably even remain essential for transcriptions to reflect the speech signal as closely as possible. For such purposes, MPTs should definitely be preferred over canonical representations because

canonical representations cannot (or only partially) represent the pronunciation variation observed in everyday speech.

A more important conclusion, however, is that the traditional validation method assigned a much higher validity rating to the MPTs than to the canonical representations. This was not confirmed by the outcome of our recognition experiment; the use of the canonical representations yielded similar recognition results. Considering the fact that the generation of MPTs is known to be time-consuming, expensive and error-prone (Cucchiarini, 1993), a preference for canonical representations seems more justified for our development task.

To conclude, we found no consistent relationship between the distance of a broad phonetic transcription to a reference transcription on the one hand, and the influence of that transcription on the recognition performance of a continuous speech recogniser on the other hand. This outcome has two implications. First of all, it suggests that ASR developers can save themselves the time and effort of collecting expensive reference transcriptions in order to validate phonetic transcriptions of speech databases or spoken language corpora. Second, and most importantly, it implies that phonetic transcriptions should preferably be validated in terms of the application they will serve because a higher resemblance to a purpose-independent reference transcription proved no guarantee for a transcription to be better suited for ASR development.

Appendix 2.1: Feature matrix to align two phonetic transcriptions of speech.

Appendix 2.1a: Articulatory feature values for the consonants.

consonant	place	voice	nasal	stop	glide	lateral	fricative	trill
p	5,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0
b	5,0	2,0	0,0	0,5	0,0	0,0	0,0	0,0
t	4,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0
d	4,0	2,0	0,0	0,5	0,0	0,0	0,0	0,0
k	2,0	1,0	0,0	0,5	0,0	0,0	0,0	0,0
f	5,0	1,0	0,0	0,0	0,0	0,0	0,5	0,0
v	5,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0
s	4,0	1,0	0,0	0,0	0,0	0,0	0,5	0,0
z	4,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0
x	2,0	1,0	0,0	0,0	0,0	0,0	0,5	0,0
G	2,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0
m	5,0	2,0	0,5	0,0	0,0	0,0	0,0	0,0
n	4,0	2,0	0,5	0,0	0,0	0,0	0,0	0,0
N	2,0	2,0	0,5	0,0	0,0	0,0	0,0	0,0
l	4,0	2,0	0,0	0,0	0,0	0,5	0,0	0,0
r	3,0	2,0	0,0	0,0	0,0	0,0	0,0	0,5
w	5,0	2,0	0,0	0,0	0,5	0,0	0,0	0,0
j	3,0	2,0	0,0	0,0	0,5	0,0	0,0	0,0
h	1,0	2,0	0,0	0,0	0,0	0,0	0,5	0,0

Appendix 2.1b: Articulatory feature values for the vowels.

vowel	length	place	tongue	round	diphthong
i	1,5	3,0	4,0	1,0	1,0
I	1,0	2,5	3,5	1,0	1,0
e	2,0	3,0	3,0	1,0	1,5
@+	2,0	3,0	3,0	2,0	1,5
E	1,0	3,0	2,0	1,0	1,0
a	2,0	2,0	1,0	1,5	1,0
A	1,0	1,0	1,5	1,5	1,0
o	2,0	1,0	3,0	2,0	1,5
O	1,0	1,0	2,0	2,0	1,0
u	1,5	1,0	4,0	2,0	1,0
y	1,5	3,0	4,0	2,0	1,0
Y	1,0	2,5	3,5	2,0	1,0
@	1,0	2,0	2,5	1,5	1,0
E+	2,0	2,5	3,0	1,0	2,0
Y+	2,0	2,5	3,0	1,0	2,0
A+	2,0	1,5	3,0	2,0	2,0

Appendix 2.2: Phone mapping 46 CGN phone set to 39 phone set.

class	example	CGN-symbol	Can/MPT symbol(s)
plosives	put	p	p
	bad	b	b
	tak	t	t
	dak	d	d
	kat	k	k
	goal	g	k
fricatives	fiets	f	f
	vat	v	v
	sap	s	s
	zat	z	z
	sjaal	S	S
	ravage	Z	z+j
	licht	x	x
	regen	G	G
	geheel	h	h
sonorants	lang	N	N
	mat	m	m
	nat	n	n
	oranje	J	n+j
	lat	l	l
	rat	r	r
	wat	w	w
	jas	j	j
short vowels	lip	I	I
	leg	E	E
	lat	A	A
	bom	O	O
	put	Y	Y
long vowels	liep	i	i
	buur	y	y
	leeg	e	e
	deuk	2	@+
	laat	a	a
	boom	o	o
	boek	u	u
schwa	gelijk	@	@
diphthongs	wijs	E+	E+
	huis	Y+	Y+
	koud	A+	A+
loan vowels	scène	E:	E
	freule	Y:	Y
	zone	O:	O
nasalised vowels	vaccin	E~	E
	croissant	A~	A
	congé	O~	O
	parfum	Y~	Y
long silence			sil
optional short silence			sp

CHAPTER 3

AUTOMATIC PHONETIC TRANSCRIPTION OF LARGE SPEECH CORPORA

Reformatted from:

Van Bael, C., Boves, L., Heuvel, H. van den, Strik, H. (2007). Automatic Phonetic Transcription of Large Speech Corpora. In: *Computer Speech and Language*, Vol. 21, pp. 652-668.

Abstract

Most large speech corpora are delivered with a lexicon that contains a canonical transcription of every word in the orthographic transcription. Such a lexicon can be used for generating a hypothetical 'canonical' phonetic transcription from the orthography. In addition, time and money permitting, some speech corpora are provided with a manually verified broad phonetic transcription of at least part of the material. Since the manual verification of phonetic transcriptions is time-consuming and expensive, we investigated whether existing automatic transcription procedures and combinations of such procedures can offer a quick and cheap alternative for the generation of phonetic transcriptions like the manually verified transcriptions delivered with large speech corpora. In our study, we used ten automatic transcription procedures to generate a broad phonetic transcription of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The performance was assessed in terms of the number and the nature of the discrepancies between the emerging phonetic transcriptions and the corresponding manually verified phonetic transcriptions delivered with the Spoken Dutch Corpus. Some of the resulting automatic transcriptions appeared to be comparable to the manually verified transcriptions.

Keywords: *Large speech corpora, Automatic phonetic transcription, Transcription evaluation.*

3.1 Introduction

In the last fifteen years we have witnessed the development of various large speech corpora. Well-known examples are TIMIT (1990), Switchboard (Godfrey et al., 1992), Verbmobil (Hess et al., 1995), the Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN, Oostdijk, 2002) and the Corpus of Spontaneous Japanese (Maekawa, 2003). The usability of such corpora largely depends on the availability of accurate annotations. It is probably fair to say that the lasting popularity of the not-so-big TIMIT corpus is due to the fact that it comes with very accurate phonetic transcriptions. Since broad phonetic transcriptions are often used and sometimes even required for diverse purposes such as lexical pronunciation variation modelling for automatic speech recognition (ASR - Strik, 2001), unit selection for speech synthesis (Mizutani and Kagoshima, 2005), automatic pronunciation training and assessment in Computer Assisted Language Learning (Neri et al., 2006; 2007) and general research on pronunciation variation (Riley et al., 1999), contemporary speech corpora are usually provided with a broad phonetic transcription of at least part of their material.

Almost all large speech corpora are provided with a phonemic lexicon that can be used to generate a hypothetical canonical phonetic representation of the material. In addition, time and money permitting, contemporary speech corpora are at least partially enriched with broad

phonetic transcriptions with the help of human transcribers in order to ensure a more accurate representation of the material. Since the employment of human transcribers is known to be exceedingly time-consuming and expensive when they have to transcribe speech from scratch, it is common practice to provide human transcribers with an example transcription they have to verify on the basis of their own perception of the speech signal. Switchboard, Verbmobil and the Spoken Dutch Corpus are three corpora which, in addition to a canonical transcription of all their material, received a manually verified phonetic transcription of a limited subset of their data (Greenberg et al., 1996; Geumann et al., 1997; Goddijn and Binnenpoorte, 2003). The example transcription the transcribers of the Spoken Dutch Corpus were presented with already reflected the obligatory cross-word assimilation and degemination processes of Dutch (Binnenpoorte and Cucchiarini, 2003). The modelling of these processes decreased the discrepancies between the original canonical example transcription and the actual speech signal, and as such it also reduced the number of required corrections and the time it took the transcribers to complete the transcription task. Although verifying automatic transcriptions is quicker and therefore also less expensive than making transcriptions from scratch, however, verification procedures also have their drawbacks.

It has been suggested that verifying example transcriptions may bias the resulting transcriptions towards the example transcriptions they are based upon (Binnenpoorte, 2006). In addition, the remaining costs are often still quite substantial. Demuynck et al. (2002) reported that their students needed 15 minutes to manually verify the transcription of one minute of public lectures, and 40 minutes for one minute of spontaneous speech. This explains why human transcribers verified an example transcription of ‘only’ one million words of the 9-million-word Spoken Dutch Corpus, and why ‘only’ four hours of Switchboard speech were phonetically transcribed as an afterthought. Still, despite these drawbacks, manually verified phonetic transcriptions are presently considered to be the best transcriptions one can feasibly obtain if large amounts of speech have to be transcribed. It is therefore worthwhile investigating whether the same transcription quality can be obtained by means of quicker and cheaper automatic transcription procedures. Because of their high transcription speed and their limited costs, automatic transcription procedures not only hold the promise of increasing transcription speed and reducing transcription costs, they even have the potential of transcribing corpora that are too large to be ever transcribed with the help of human transcribers.

Several studies already reported benefits of using automatic phonetic transcriptions (APTs) for the development of ASR systems (e.g. Riley et al., 1999; Saraçlar and Khundanpur, 2004; Tjalve and Huckvale, 2005; Wester, 2003; Yang and Martens, 2000) and speech synthesis systems (e.g. Bellegarda, 2005; Jande, 2005; Wang et al. 2005). However, since in these studies the transcriptions were used as mere tools for the development of specific speech applications, the procedures with which the transcriptions were generated were not evaluated in terms of their ability to approximate the quality of manually verified phonetic transcriptions. Therefore, our study was aimed at investigating whether existing automatic transcription procedures and combinations of such procedures can approximate manually verified phonetic transcriptions and, consequently, whether they can offer a sound

alternative to commonly used but nonetheless time-consuming and expensive verification procedures for the transcription of large speech corpora.

We assessed the quality of ten transcription procedures in terms of the resemblance of their transcriptions and a manually verified broad phonetic transcription of read speech and of spontaneous telephone dialogues from the Spoken Dutch Corpus. Since we aimed at approximating transcriptions that were made with a limited symbol set and that originated from canonical example transcriptions, it should be clear that our experiments were not aimed at comparing or improving the transcription procedures in terms of the accuracy with which they can describe the actual speech signal.

In order to ensure the applicability of the transcription procedures in contexts where only minimal resources are available, we optimised our procedures with limited resources and minimal human effort. Most procedures only required a standard continuous speech recogniser, an algorithm to align phonetic transcriptions, an orthographically transcribed corpus, a canonical lexicon and a manually verified phonetic transcription of a relatively small sample of the corpus. The manually verified phonetic transcription was required to tune the transcription procedures and to evaluate their performance. Some procedures also required software for the implementation of decision trees, and some (also) a list of phonological processes describing pronunciation variation in the language under investigation (Dutch). Expert human effort was limited to the compilation of such a list of phonological processes, and the aforementioned manual verification of an example transcription of a limited amount of speech.

This paper is organised as follows. In Section 3.2, we introduce the material and tools we used in our study. Section 3.3 sketches the various transcription procedures. Section 3.4 presents the evaluation of the emerging transcriptions. In Section 3.5, we discuss our results, and in Section 3.6, we formulate our conclusions.

3.2 Material and tools

3.2.1 Speech material

We worked with speech material from the Spoken Dutch Corpus (Oostdijk, 2002). We considered speech of native speakers from the Netherlands only. In order not to base the assessment of the transcription procedures on the transcription of speech from one particular speech style, we chose to work with read speech as well as spontaneous telephone dialogues.

The read speech was recorded at 16 kHz (16-bit PCM) with high-quality table-top microphones for the compilation of a library for the blind. The telephone dialogues, comprising much more spontaneous speech, were recorded at 8 kHz (8-bit A-law) through a telephone platform. As part of the orthographic transcription process, the speech material was manually segmented into speech chunks of approximately 3 seconds each. The transcribers were instructed to put chunk boundaries in naturally occurring pauses. Only if speech

stretched for substantially longer than 3 seconds without a silent pause, the transcribers were requested to put chunk boundaries between adjoining words with minimal cross-word co-articulation. We adhered to this chunk-level annotation. In order to be able to focus on phonetic transcription proper, we excluded speech chunks that, according to the orthographic transcription, contained non-speech, unintelligible speech, broken words and foreign speech. Chunks containing overlapping speech (in the telephone dialogues) were excluded as well.

The statistics of the data are presented in Table 3.1. We divided the data of each speech style into a training set, a development set and an evaluation set. To this end, we listed all speech chunks of all speakers, we randomised their ordering, and we extracted the subsets. This guaranteed mutually exclusive data sets with similar material. The resulting data sets of the two speech styles differ in size, but we preferred to work with all the material meeting our requirements rather than ignoring half of the read speech.

Table 3.1: Statistics of the data sets.

speech style		training sets	development sets	evaluation sets
read speech	# word tokens	532,451	7,940	7,940
	hh:mm:ss	44:55:59	0:40:10	0:41:39
	# distinct speakers	561	126	126
telephone dialogues	# word tokens	263,501	6,953	6,955
	hh:mm:ss	18:20:05	0:30:02	0:29:50
	# distinct speakers	344	92	91

3.2.2 Canonical lexicon

Our canonical lexicon was a comprehensive multi-purpose in-house lexicon. It was compiled by merging various existing lexical resources such as CELEX (Baayen et al, 1995), RBN (ReferentieBestand Nederlands, 2005) and PAROLE (PAROLE lexicon, 2005). The pronunciation forms in the lexicon reflected the standard pronunciation of words as they would be carefully pronounced in isolation according to the obligatory word-internal phonological processes of Dutch (Booij, 1999). Each word was represented by just one standard broad phonetic transcription. We ignored all information about syllabification and syllabic stress in order to ensure the applicability of the transcription procedures in research contexts where a lexicon with this kind of linguistic information is unavailable.

3.2.3 Reference transcriptions (RTs)

We used the manually verified phonetic transcriptions of the Spoken Dutch Corpus as Reference Transcriptions (RTs) to tune (with the RTs of the development sets) and evaluate (by means of the RTs of the evaluation sets) the transcription procedures. The manually verified transcriptions of the Spoken Dutch Corpus were generated in three steps. First, the canonical representation of every word was selected from the lexicon. Subsequently, two cross-word phonological processes of Dutch, voice assimilation and degemination, were applied to the phones at word boundaries in order to decrease the discrepancies between the canonical transcription and the speech signal. The resulting transcriptions were finally verified and corrected by human transcribers. The transcribers acted according to a strict protocol instructing them to change the example transcription only if they were certain that it did not correspond to the speech signal (Binnenpoorte and Cucchiarini, 2003).

3.2.4 Continuous speech recogniser (CSR)

Except for the canonical transcriptions, all automatic phonetic transcriptions (APTs) were generated by means of a continuous speech recogniser (CSR) that was based on Hidden Markov Models and that was implemented with the HTK Toolkit (Young et al., 2001). Our CSR used 39 gender- and context independent, but speech style-specific acoustic models with 128 Gaussian mixture components per state (37 phone models, one model for silences of 30 ms or more and one model for the optional silence between words).

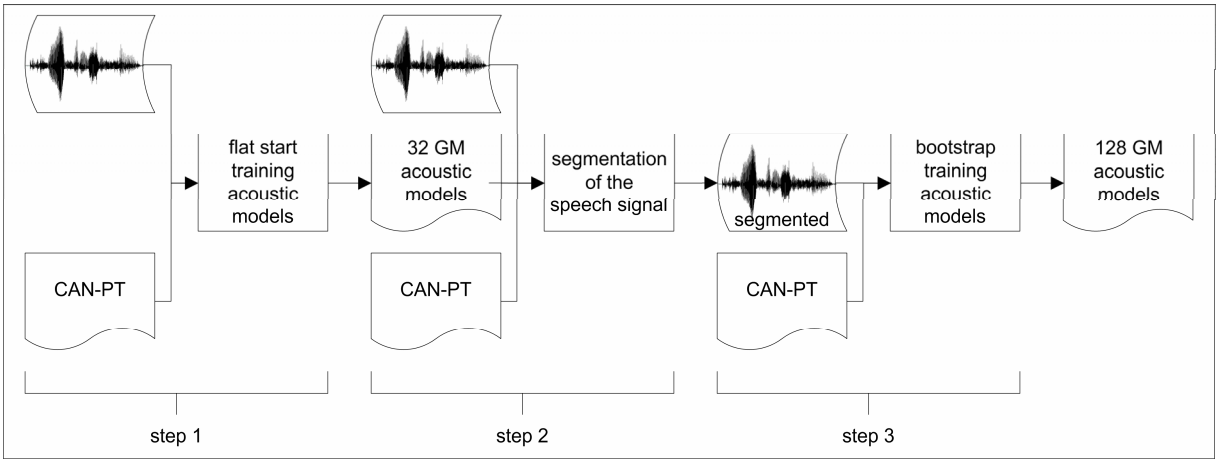

Figure 3.1: The procedure by means of which the acoustic models were trained.

We trained our acoustic models in three stages with the canonical transcriptions (CAN-PTs) of the training data (see Figure 3.1). First, we trained flat start acoustic models with 32 Gaussian mixture components in 41 iterations. Subsequently, we used these models to obtain a more realistic segmentation of the speech material. We used this segmentation to bootstrap a new set of acoustic models, which we retrained (with 55 iterations) to models with 128 Gaussian mixture components per state. Experiments with the development sets of both speech styles showed that acoustic models with 128 mixture components yielded transcriptions that resembled the target transcriptions more closely than transcriptions that were generated with models with fewer mixture components per state.

3.2.5 Algorithm for Dynamic Alignment of Phonetic Transcriptions (ADAPT)

ADAPT (Elffers et al., 2005) is a dynamic programming algorithm designed to align two strings of phonetic symbols according to the articulatory distance between them. We used ADAPT to align phonetic transcriptions for the generation of lexical pronunciation variants for forced recognition (Section 3.3.1), and for the quality assessment of the automatic phonetic transcriptions through their alignment with the manually verified reference transcriptions (Section 3.3.2).

3.3 Method

We investigated the suitability of ten automatic transcription procedures for the phonetic transcription of large speech corpora. The transcription procedures are introduced in Section 3.3.1. In Section 3.3.2 we describe the evaluation procedure by means of which the automatic phonetic transcriptions and, consequently, the transcription procedures were assessed.

3.3.1 Generation of phonetic transcriptions with different transcription procedures

Figure 3.2 shows the ten transcription procedures by means of which our APTs were generated. We used three generic procedures, two combinations of these procedures, and five procedures in which we used decision trees to further tune the output of the aforementioned procedures towards the type of transcription we were trying to approximate. Most of the procedures required the tuning of several parameters to optimally approximate the RTs of the data in the development sets. The optimal parameter settings were subsequently used for the transcription of the data in the evaluation sets.

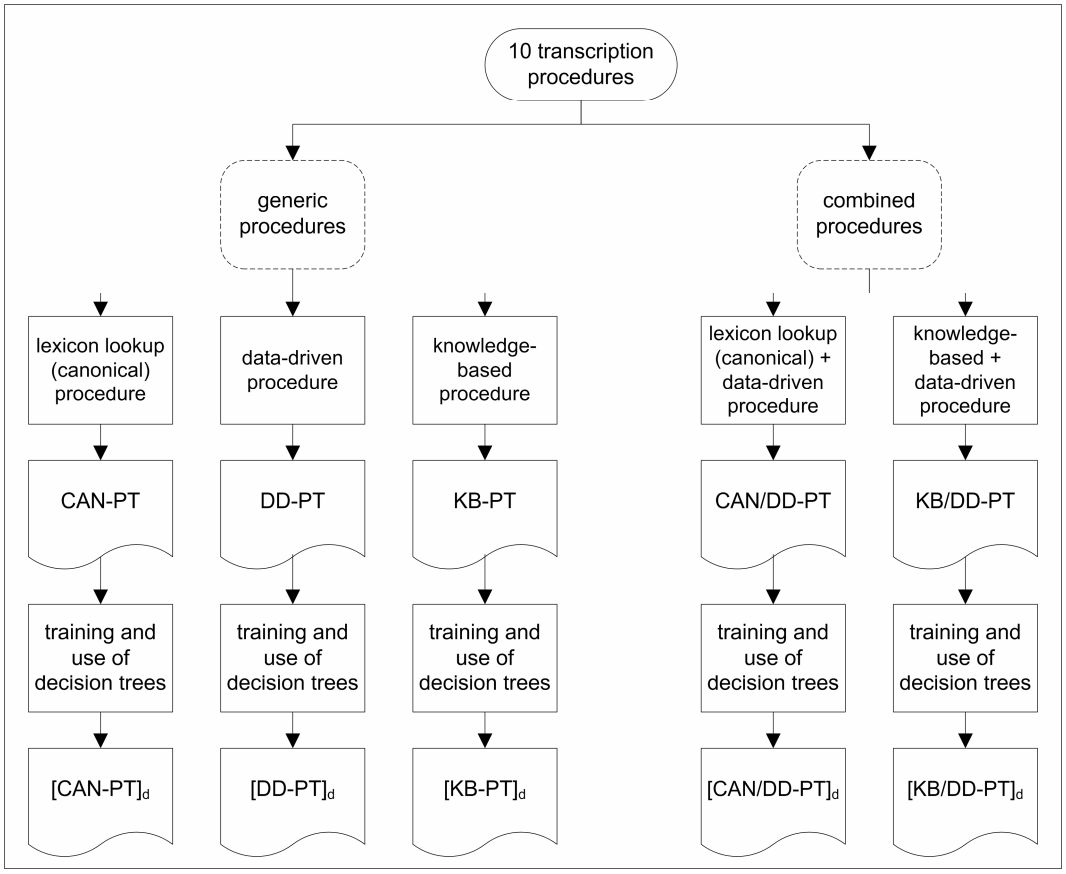

Figure 3.2: Overview of the ten investigated transcription procedures.

Generic transcription procedures

Lexicon lookup (canonical) transcription procedure

The canonical phonetic transcriptions (CAN-PTs) were generated through a lexicon lookup procedure. Cross-word processes were not modelled. In general, canonical transcriptions like these can be easily obtained, since many corpora are provided with an orthographic transcription and a canonical pronunciation lexicon comprising a broad phonetic transcription of the words in the orthographic transcription.

Data-driven transcription procedure

The data-driven phonetic transcriptions (DD-PTs) were based on the acoustic *data*. The DD-PTs were generated through constrained phone recognition; a CSR segmented and labelled the speech signal by means of its acoustic models and a phonotactic model. The phonotactic models (one for each speech style) were trained on the RTs of the development data.

Figure 3.3 shows the last three steps of the data-driven transcription procedure. The first step, the training of the phonotactic models, is not included in the Figure. We trained bigram, trigram, four-gram, five-gram and six-gram models. Since the current version of HTK (v.3.2) only supports the use of unigram and bigram models in its first decoding pass, we used a bigram model in the first pass, and higher order n-gram models to rescore the resulting phone lattices (step 3). The final phonetic transcription of the data (step 4) was obtained with a four-gram model. Transcription experiments with the development data of both speech styles indicated that the use of four-gram models yielded transcriptions that resembled the RTs more closely than the bi-, tri-, penta- and hexagram phonotactic models.

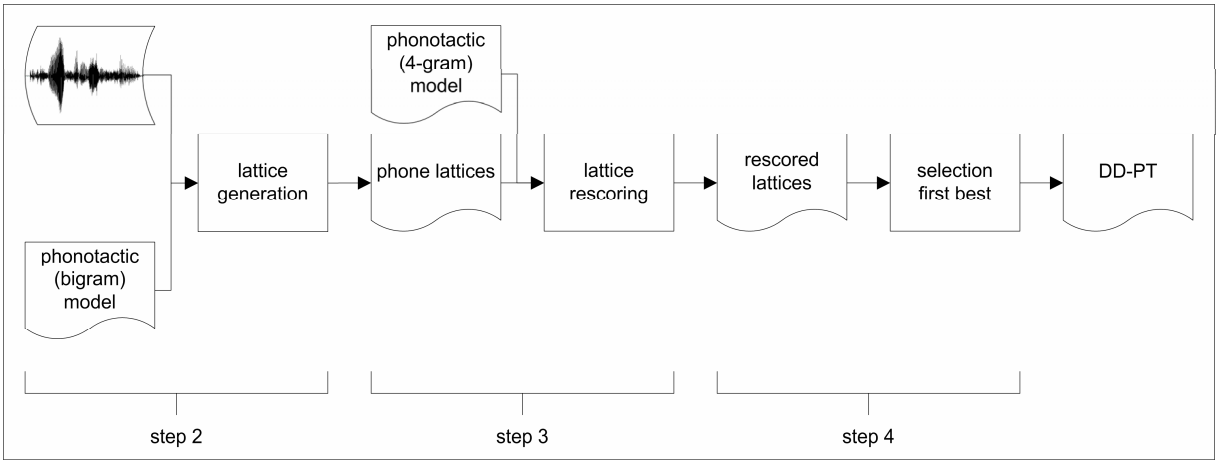

Figure 3.3: Data-driven phonetic transcription through constrained phone recognition (step 1 – the training of the phonotactic models – is not included).

Knowledge-based transcription procedure

ASR research often draws on the linguistic literature for the extraction of knowledge to generate lexical pronunciation variants for recognition (Kessens et al., 1999; Strik, 2001). Figure 3.4 illustrates the three-step procedure we used to generate knowledge-based phonetic transcriptions (KB-PTs).

We first compiled a list of 20 prominent phonological processes from the literature on the phonology of Dutch (Booij, 1999). We implemented these processes as context-dependent rewrite rules modelling both within-word and cross-word contexts in which phones from the CAN-PT could be deleted, inserted or substituted with other phones. Most of the processes identified by Booij (1999) could be described in terms of operations on phoneme symbols or articulatory features. However, some of the processes could only be described with information about the prosodic or syllabic structure of words. We reformulated most of these processes in terms of phonetic symbols and features, since we wanted to exclude non-segmental information from our experiments (see Section 3.2.2). We implemented the rules in a conservative manner in order to minimise the risk of over-generation. The resulting rule set

comprised phonological rules describing progressive and regressive voice assimilation, nasal assimilation, t-deletion, n-deletion, r-deletion, schwa deletion, schwa epenthesis, palatalisation, degemination and more specific rules modelling pronunciation variation in high-frequency words (e.g. demonstratives) in Dutch. The reduction and the deletion of full vowels, two prominent phonological processes in Dutch, were not implemented since they could not be formulated without the explicit use of supra-segmental information.

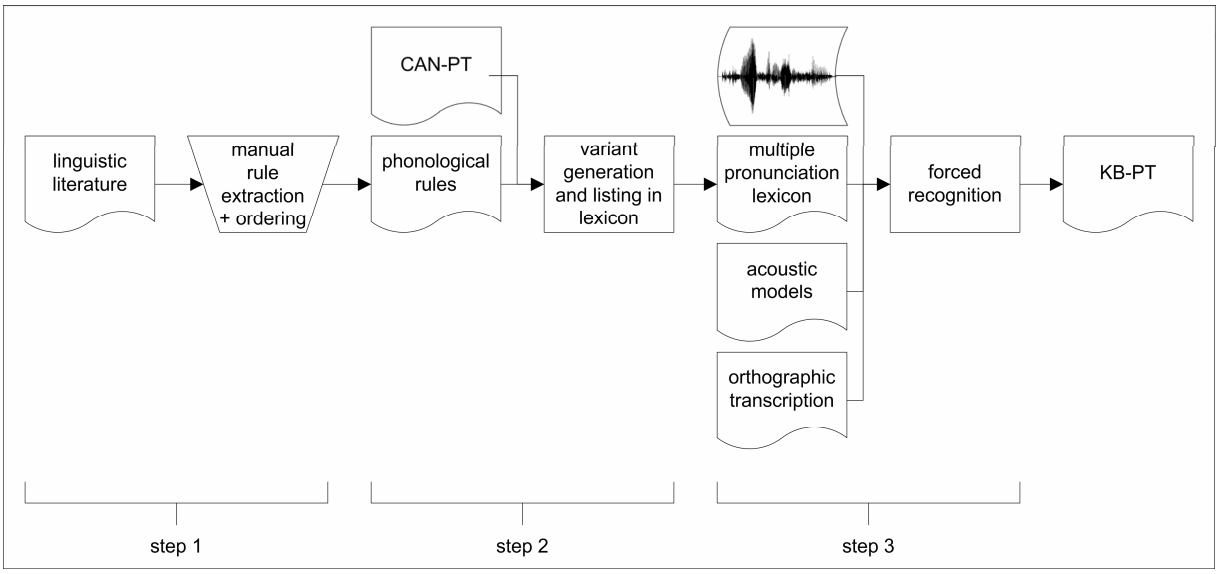

Figure 3.4: Knowledge-based phonetic transcription.

In the second step of the procedure, we used the phonological rewrite rules to generate pronunciation variants from the CAN-PTs of the speech chunks. Note that it was necessary to apply the rules to the speech chunks rather than to the words in isolation, for cross-word processes could only be modelled if the neighbouring words were known. The rules only applied once, and their order of application was manually optimised. Analysis of the resulting pronunciation variants suggested that hardly any implausible variants were generated, and that no obvious variants were missing. It may well be, however, that two-level rules (Koskenniemi, 1983) or an iterative application of rewrite rules are needed for the generation of all plausible pronunciation variants in languages other than Dutch.

In the third step of the procedure, the pronunciation variants (including the original CAN-PTs) of each individual speech chunk were listed. Since the linguistic literature hardly ever provides accurate information on the frequency of phonological processes, and since trustworthy priors can only be learned from the analysis of a sufficiently large amount of manually verified transcriptions (the amount of manual transcriptions that is hardly every available), our knowledge-based pronunciation variants did not comprise prior probabilities. The optimal knowledge-based phonetic transcription (KB-PT) was identified through forced recognition.

Combinations of generic transcription procedures

After having generated the CAN-PTs, DD-PTs and KB-PTs, we combined these transcriptions to obtain new transcriptions. Chunk-level pronunciation variants were generated through the automatic alignment of two APTs at a time. Since the KB-PTs were based on the CAN-PTs, we only combined the CAN-PTs with the DD-PTs (CAN/DD-PT) and the KB-PTs with the DD-PTs (KB/DD-PT) to generate new pronunciation variants in addition to the original CAN-PTs, DD-PTs and KB-PTs. Figure 3.5 shows how new pronunciation variants were generated through the alignment of the phones in two different APTs. These pronunciation variants were listed, after which our CSR was forced to choose the best matching pronunciation variant for every chunk of words in the orthographic transcriptions. The three steps of this combined transcription procedure are illustrated in Figure 3.6.

CAN-PT	A n	A t	d @	A p @ l t a r t
+				
DD-PT	A n	A t	d	A b @ l t a t
multiple pronunciation variants	A n	A t	d @	A p @ l t a r t
	A n	A t	d	A p @ l t a r t
	A n	A t	d @	A b @ l t a r t
	A n	A t	d	A b @ l t a r t
	A n	A t	d @	A p @ l t a t
	A n	A t	d	A p @ l t a t
	A n	A t	d @	A b @ l t a t
	A n	A t	d	A b @ l t a t

Figure 3.5: Generation of pronunciation variants through the alignment of two phonetic transcriptions.

We combined the APTs from the different transcription procedures to provide our CSR with additional linguistically plausible pronunciation variants for the words in the orthographic transcriptions. After all, canonical transcriptions do not model pronunciation variation, and our KB-PTs only modelled the pronunciation variation that was manually implemented in the form of phonological rewrite rules. The DD-PTs, however, were based on the speech signal. Therefore, they were potentially better at representing the actual speech signal, at the risk of being linguistically less plausible than the CAN-PTs and the KB-PTs. It was reasonable to expect that the combination of the different transcription procedures would reinforce the advantages and alleviate the disadvantages of the individual procedures.

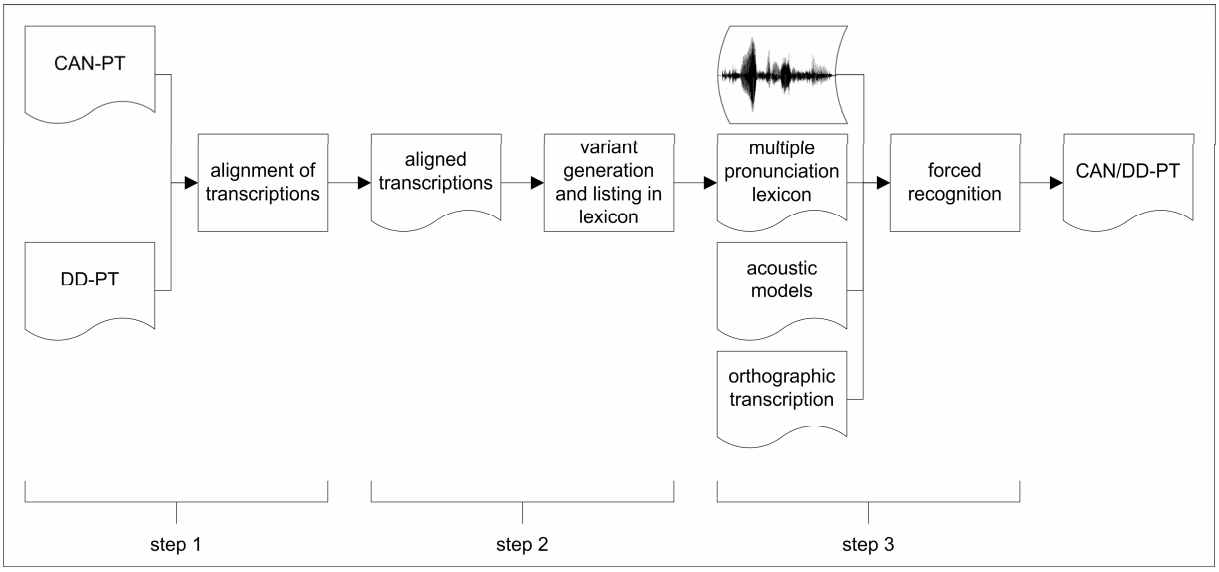

Figure 3.6: Combination of transcription procedures (in this case: CAN-PT and DD-PT).

Transcription procedures with decision trees

The use of data-driven transcription procedures can result in too many, too few or very unlikely lexical pronunciation variants (Wester, 2003). Therefore, ASR developers often use decision trees to reduce the number of unlikely pronunciation variants and to optimise the number of plausible pronunciations in recognition lexica (Riley et al., 1999; Wester, 2003). Figure 3.7 illustrates our four-step procedure to improve the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs through the use of decision tree filtering. The decision trees were generated with the C4.5 algorithm (Quinlan, 1993), which is provided with the Weka package (Witten and Frank, 2005), a collection of Java-based machine learning algorithms.

First, the APT (each of the aforementioned transcriptions individually) and the RT of the development data were aligned. Second, all the phones and their context phones in the APT were listed. We will call these phone sequences 'phonetic windows' for the sake of convenience. The size of these phonetic windows was limited to the target phone and its immediately left and right neighbours. Word boundaries were included as extra information in order to model pronunciation variation across word boundaries. The correspondences of the phonetic windows in the APT and the phones in the RT, and the frequencies of these correspondences were used to estimate:

$$P(RT\_phone \mid APT\_phonetic\_window) \quad (1)$$

i.e. the probability of a phone in the reference transcription given a particular phonetic window in the APT.

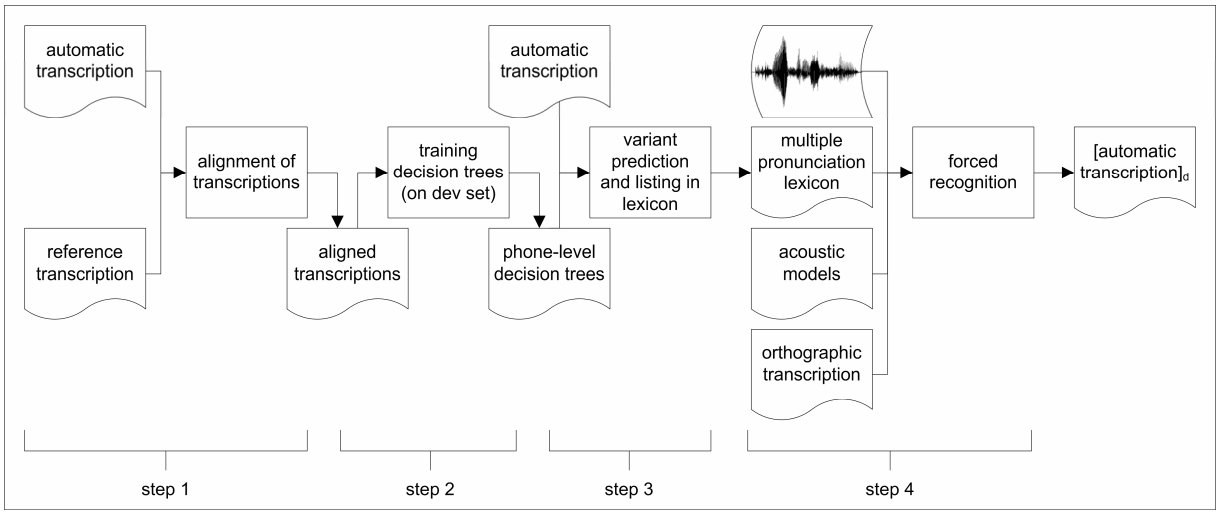

Figure 3.7: Automatic phonetic transcription with decision trees.

Figure 3.8 shows a simplified version of the decision tree trained for the phone /e/. The tree strongly predicts (because in the development data, in 12 out of 13 cases, it was the case) that a word-initial /e/ or an /e/ preceded by /@/ and followed by either /@/, /n/ or /j/ should be deleted. Based on the observations in the development data, in all other contexts, all /e/s in the APT should remain. The application of this knowledge is illustrated in the lower half of Figure 3.8.

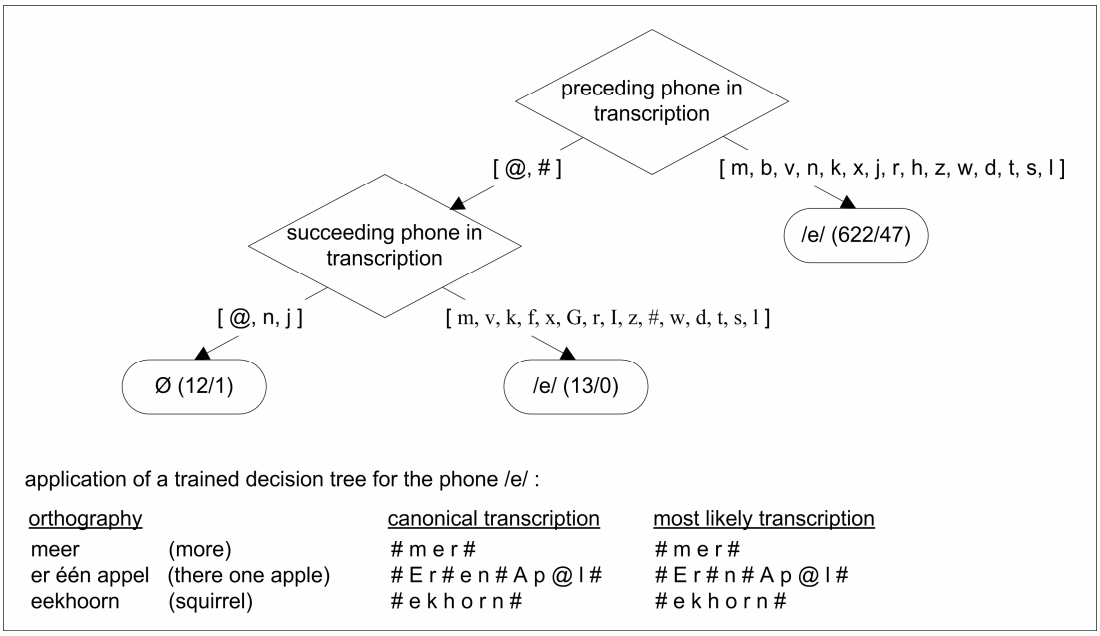

Figure 3.8: Illustration and application of a decision tree for the phone /e/ given its left and right context phones (# = word boundary, Ø = deletion of /e/). This Figure is based on Figure 2 in Wester (2002:108).

In the third step of the procedure, the decision trees were used to generate plausible pronunciation variants for the APT of the unseen evaluation data. The decision trees were used to predict:

$$P\ (pronunciation\_variants \mid APT\_phonetic\_window) \quad (2)$$

i.e. the probability of a phone with optional pronunciation variants given a particular phonetic window in the APT. In our experiments, all phone variants with a probability lower than 0.1 were ignored (Wester, 2003). This reduced the number of pronunciation variants and, more importantly, it pruned unlikely pronunciation variants originating from idiosyncrasies in the original APT. The retained phone-level variants were combined to word-level variants. These variants were listed in a multiple pronunciation lexicon. Their probabilities were normalised so that the probabilities of all variants of a word added up to 1.

In the fourth and final step of the transcription procedure, our CSR selected the most likely pronunciation variant for every word in the orthographic transcription. The consecutive application of the decision trees to the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs resulted in new transcriptions hereafter referred to as $[\text{CAN-PTs}]_d$, $[\text{DD-PTs}]_d$, $[\text{KB-PTs}]_d$, $[\text{CAN/DD-PTs}]_d$ and $[\text{KB/DD-PTs}]_d$.

3.3.2 Evaluation of the phonetic transcriptions and the transcription procedures

The APTs of the data in the evaluation sets were evaluated in terms of their deviations from the manually verified RTs. We compared the transcriptions by means of ADAPT (Elffers et al., 2005). The disagreement metric was defined as:

$$Percentage\ \ disagreement = \left(\frac{Sub_{phone} + Del_{phone} + Ins_{phone}}{N_{phone}} \right) \times 100\% \quad (3)$$

i.e. the sum of all phone substitutions (Sub_{phone}), deletions (Del_{phone}) and insertions (Ins_{phone}) divided by the total number of phones in the RT (N_{phone}). Considering the aim of our research, a smaller deviation from the reference transcription indicated a 'better' transcription. A detailed analysis of the number and the nature of the deviations allowed us to systematically investigate the magnitude and the nature of the improvements and deteriorations caused by the use of the different transcription procedures.

3.4 Results

The figures in Table 3.2 show the disagreements between the APTs and the RTs of the evaluation data. From top to bottom and from left to right we see the disagreement scores (%dis) between the different APTs and the RTs of the read speech and the telephone

dialogues. In addition, the statistics of the phone substitutions (subs), deletions (dels) and insertions (ins) are presented in order to provide insight into the nature of the disagreements.

Table 3.2: Evaluation of the transcription procedures through a comparison of the transcriptions with reference transcriptions. Fewer disagreements (%dis) indicate better transcriptions and therefore better transcription procedures.

	read speech				telephone dialogues			
	subs	dels	ins	%dis	subs	dels	ins	%dis
CAN-PT	6.3	1.2	2.6	10.1	9.1	1.1	8.1	18.3
DD-PT	16.1	7.4	3.6	27.0	26.0	18.0	3.8	47.8
KB-PT	6.3	3.1	1.5	10.9	9.0	2.5	5.8	17.3
CAN/DD-PT	13.1	2.0	4.8	19.9	21.5	6.2	7.1	34.7
KB/ DD-PT	12.8	3.1	3.6	19.5	20.5	7.8	5.4	33.7
$[\text{CAN-PT}]_d$	4.8	1.6	1.7	8.1	7.1	3.3	4.2	14.6
$[\text{DD-PT}]_d$	15.7	7.4	3.5	26.7	26.0	18.6	3.8	48.3
$[\text{KB-PT}]_d$	5.0	3.2	1.2	9.4	7.1	3.5	4.2	14.8
$[\text{CAN/DD-PT}]_d$	12.0	2.3	4.3	18.5	20.1	7.2	5.5	32.8
$[\text{KB/ DD-PT}]_d$	11.6	3.1	3.1	17.8	19.3	9.4	4.5	33.1

The proportions of disagreements observed in the CAN-PTs and the KB-PTs differed significantly from each other for both speech styles ($p < .01$; we report *t*-tests throughout this article). However, the CAN-PT of the read speech was more similar to the RT than the KB-PT ($\Delta = 6.3\%$ rel.), while the opposite held for the telephone dialogues ($\Delta = 5.9\%$ rel.). In both speech styles, the proportion of substitutions was about equal in the CAN-PT and the KB-PT. Deletions made up only a very small proportion of the discrepancies, so the most important difference was in the insertions; the proportion of insertions was much higher in the telephone speech than in the read speech. The ten most frequent mismatches in the CAN-PTs and the KB-PTs of the two speech styles are presented in Tables 3.3 and 3.4, respectively. We observed many similar mismatches due to voiced/unvoiced classification of obstruents, as well as insertions of schwa and various consonants (in particular /r/, /t/ and /n/). Most substitutions and deletions (about 62-75% for the various transcriptions) occurred at word boundaries, but the absolute numbers in the KB-PTs were lower due to the cross-word pronunciation modelling inherent to the knowledge-based transcription procedure.

The disagreement scores obtained with the DD-PTs were much higher than the scores obtained with the CAN-PTs and the KB-PTs. This holds for both speech styles. Most discrepancies between the DD-PTs and the RTs were deletions and (a variety of) substitutions. In addition to consonant substitutions due to voicing, we observed various consonant substitutions due to place of articulation, and vowel substitutions with schwa (and vice versa).

Table 3.3: 10 most frequent mismatches between the CAN-PTs and the RTs.

read speech						telephone dialogues					
substitutions		deletions		insertions		substitutions		deletions		insertions	
RT	CAN-PT	RT	CAN-PT	RT	CAN-PT	RT	CAN-PT	RT	CAN-PT	RT	CAN-PT
f	v			-	@	f	v	n	-	-	r
s	z			-	r	s	z			-	h
d	t			-	t	@	E			-	n
x	G			-	n	d	t			-	t
g	k			-	d	x	G				

Table 3.4: 10 most frequent mismatches between the KB-PTs and the RTs.

read speech						telephone dialogues					
substitutions		deletions		insertions		substitutions		deletions		insertions	
RT	KB-PT	RT	KB-PT	RT	KB-PT	RT	KB-PT	RT	KB-PT	RT	KB-PT
f	v	@	-	-	h	f	v	@	-	-	@
s	z	n	-			s	z			-	r
@	E	r	-			d	t			-	t
x	G					x	G			-	d
d	t									-	n
t	d										

The proportions of disagreements in the CAN/DD-PTs and the KB/DD-PTs were lower than in the DD-PTs, but much higher than in the CAN-PTs and KB-PTs. Thus, the combination of the transcription procedures improved the DD-PTs, but deteriorated the CAN-PTs and KB-PTs. The CAN/DD-PTs and the KB/DD-PTs comprised twice as many substitutions as the CAN-PTs and the KB-PTs. Whereas the highly increased number of deletions in the CAN/DD-PT of the telephone dialogues (as compared to the CAN-PT) coincided with a - be it moderate - decrease of insertion errors, the CAN/DD-PT of the read speech showed even more insertions than the CAN-PT.

We used decision trees to narrow the gap between the ten aforementioned APTs (5 procedures x 2 speech styles) and the reference transcriptions. In nine out of ten cases, the use of decision trees improved the original transcriptions; only the $[\text{DD-PT}]_d$ of the telephone dialogues comprised more disagreements than the original DD-PT. The magnitude of the improvements differed substantially, though. The improvements were negligible for the DD-PTs, somewhat larger for the APTs that emerged from the combined procedures, and most outspoken for the CAN-PTs and the KB-PTs. This is quite remarkable, because one would expect the biggest improvement for the worst baseline. Our results show the opposite. For both speech styles, the $[\text{CAN-PT}]_d$ proved most similar to the RT. The $[\text{KB-PTs}]_d$ were slightly worse. The $[\text{CAN-PTs}]_d$ comprised on average 20.5% less mismatches with the RTs than the original CAN-PTs, which is a significant improvement at a 99% confidence level.

Likewise, we observed on average 14.1% less mismatches in the $[\text{KB-PTs}]_d$ than in the original KB-PTs ($p < .01$).

3.5 Discussion

3.5.1 Reflections on the evaluation procedure

We assessed our automatic phonetic transcriptions in terms of their resemblance to reference transcriptions that were based on example transcriptions. Previous studies have shown that the use of an example transcription for verification speeds up the transcription process (relative to manual transcription from scratch), but that it also tempts human experts into adhering to the example transcription despite contradicting acoustic cues in the speech signal. Demuynck et al. (2004), for example, reported cases where human transcribers preferred not to change the example transcription in the presence of contradicting acoustic cues, and cases where transcribers left phones in the example transcription that could not be aligned with a specific portion of the speech signal.

This observation is important for our study, because it implies that our RTs may have been biased towards the canonical example transcriptions they were based upon. Considering that both the RTs and the KB-PTs were based on the CAN-PTs, it is reasonable to assume that the quality assessments of the CAN-PTs and the KB-PTs have been positively biased in our experiments. At the same time, the assessment of the DD-PTs may have been negatively biased, since these transcriptions were only based on the signal. Most probably, the transcribers' instruction to accept the example transcription as long as the acoustic evidence did not unequivocally suggest another transcription has contributed to the discrepancies between the DD-PTs and the RTs.

3.5.2 On the suitability of a low-cost transcription procedure for the automatic phonetic transcription of large speech corpora

Generic transcription procedures

Lexicon lookup (canonical) transcription procedure

The quality of the CAN-PT of the telephone dialogues (18.3% disagreement) was rather good as compared to human inter-labeller disagreement scores reported in the literature. Greenberg et al. (1996), for example, reported 25 to 20% disagreements between human transcriptions of American English telephone conversations, and Kipp et al. (1997) reported 21.2 to 17.4% inter-labeller disagreements between human transcriptions of German spontaneous speech.

Binnenpoorte (2006), assessing the inter-labeller disagreements between manually verified phonetic transcriptions of spontaneous conversations in the Spoken Dutch Corpus, reported between 14 and 11.4% disagreements. The proportion of disagreements between the CAN-PT and the human RT (10.1% disagreement) of the read speech was still relatively high as compared to human inter-labeller disagreement scores reported in the literature. Kipp et al. (1996) reported 6.9 to 5.6% disagreements between human transcriptions of German read speech, and Binnenpoorte (2006) reported 6.2 to 3.7% inter-labeller disagreements between manually verified transcriptions of Dutch read speech.

Considering the very low cost of CAN-PTs, and considering the similarities with previously published human inter-labeller disagreement scores, it appears that the production of CAN-PTs is a viable option in transcription projects in which limited resources are available. However, we still found a high proportion of substitutions and insertions at word boundaries. This is not surprising, because cross-word phonological processes were not accounted for in the CAN-PTs.

Data-driven transcription procedure

Constrained phone recognition proved suboptimal to approximate the manually verified phonetic transcriptions. The high number and the wide variety of substitutions suggest that the use of phonotactic models did not sufficiently tune our CSR towards the RTs. The high number of deletions implies that, in spite of extensive tuning of the phone insertion penalty, our CSR had too large a preference for transcriptions containing fewer symbols. Close inspection of the DD-PTs suggested that many deletions were systematic, but unlikely. Thus, it is not likely that the discrepancy between the DD-PTs and the RTs are fully due to a bias towards canonical representations of the human transcribers. Kessens and Strik (2004) observed that the use of shorter acoustic models for sounds like /@/ (e.g. two-state models that can be aligned to signal segments as short as 20 ms instead of the conventional three-state models that cover at least 30 ms of the speech signal) may reduce this tendency for deletions, but the diverse nature of the deletions in our results makes a substantial reduction of deletions through the mere use of shorter acoustic models rather unlikely.

Knowledge-based transcription procedure

The use of linguistic knowledge to model pronunciation variation at the lexical level improved the quality of the transcription of the telephone dialogues, but it deteriorated the transcription of the read speech. The availability of pronunciation variants is probably more beneficial for the transcription of spontaneous speech, since more spontaneous speech is often characterised by a larger degree of pronunciation variation (Goddijn and Binnenpoorte, 2003). Most probably, the CSR often preferred non-canonical variants for the transcription of the read speech, while the human transcribers had a preference for the canonical example transcription, according to their instruction.

The knowledge-based multiple pronunciation lexicon of the telephone dialogues comprised on average 1.39 pronunciation variants per word, the lexicon of the read speech 1.47 variants per word. The higher average number of pronunciation variants in the read speech lexicon can be explained by the fact that the pronunciation variants of both speech styles were derived from the canonical transcriptions by applying a fixed set of rules. Since the words in the telephone dialogues were shorter than the words in the read speech (an average of 3.3 vs. 4.1 canonical phones per word in the telephone dialogues and the read speech), the canonical transcription of the telephone dialogues was less susceptible to the application of rewrite rules than the canonical transcription of the read speech.

In order to estimate the possible impact of the application of knowledge-based rewrite rules on the CAN-PTs, we computed the maximum and minimum accuracy that could be obtained with the knowledge-based recognition lexica for read and spontaneous speech. For every chunk, every combination of the pronunciations of the words was aligned with the RT, and the highest and the lowest disagreement measures were retained. We found that the knowledge-based recognition lexicon of the telephone dialogues was able to provide KB-PTs of which 22.6 to only 13.2% of the phones differed from the RT. The knowledge-based lexicon of the read speech was able to provide KB-PTs of which 16.3 to only 7.4% of the phones differed from the RT. The eventual quality of the KB-PTs (17.3% and 10.9% disagreement for the telephone dialogues and the read speech, respectively) shows that there was still room for improvement; the acoustic models of our CSR often opted for suboptimal transcriptions.

Combinations of generic transcription procedures

The blend of data-driven pronunciation variants with canonical or knowledge-based variants into CAN/DD and KB/DD lexica allowed our CSR to better approximate human transcription behaviour than through constrained phone recognition alone, but the combination of the procedures did not outperform the canonical lexicon lookup (CAN-PT) and the knowledge-based transcription procedure (KB-PT). The improvement with regard to the original DD-PTs must have been due to the fact that the CSR could now only select phoneme sequences from the multiple pronunciation lexica. This constituted a substantial bias in the direction of the RTs as compared to the constrained phone recognition through which the DD-PTs were generated. The fact that the CAN/DD and KB/DD transcriptions suffered from the addition of the signal-based pronunciation variants could be due to the added variants closer resembling the signal than the canonical representations did (and the representations derived by means of phonological rules), whereas the transcribers adhered to the canonical example transcriptions. We conclude that the mere combination of signal-based and canonical or knowledge-based lexical pronunciation variants was not effective for approximating the manually verified phonetic transcriptions.

Transcription procedures with decision trees

Contrary to our expectations, the $[\text{DD-PT}]_d$ of the telephone dialogues differed more from the RT (though not significantly more, $p > .1$) than the original DD-PT. The $[\text{DD-PT}]_d$ of the read speech was only slightly (again, not significantly, $p > .1$) better than the original DD-PT. The inability of the decision trees to tune the data-driven transcriptions towards the RTs was probably due to the high degree of confusability in the recognition lexica in the absence of reliable estimates of prior probabilities. The recognition lexicon for the telephone dialogues had an average of 9.5 variants per word, and the lexicon for the read speech an average number of 3.5 variants per word.

Note that, contrary to the pronunciation variants in the knowledge-based recognition lexica, the pronunciation variants in the $[\text{DD-PT}]_d$ lexica were based on the speech signal rather than on the application of phonological rewrite rules on the CAN-PT. This resulted, in particular for the $[\text{DD-PTs}]_d$ of the more spontaneous telephone dialogues, in more discrepancies with the RTs, all of which were modelled in the decision trees. Even after pruning unlikely pronunciation variants from the decision trees, the decision trees apparently still comprised enough pronunciation variants to boost the average number of pronunciation variants per word in the recognition lexica. From experience with ASR tasks it is known that an average number of 2.5 pronunciations per word is close to the optimum in terms of word error rate (Kessens et al., 2003). It was shown that the addition of more pronunciation variants to recognition lexica increases the risk of lexical confusability. In our study, for the purpose of automatic phonetic transcription, the CSR had to choose between highly similar alternatives. Apparently, an average of 9.5 pronunciation variants per word in the recognition lexicon for the telephone dialogues was too high, whereas an average of 3.5 variants in the lexicon for the read speech seemed tolerable, even though it was more than the optimum of 2.5 variants previously reported for ASR.

The small improvements obtained through the use of decision trees for the enhancement of the CAN/DD-PTs and the KB/DD-PTs, as well as the large improvements obtained through the use of decision trees for the enhancement of the CAN-PTs and the KB-PTs can be explained along the same line of reasoning. The numerous discrepancies between the CAN/DD-PTs and the KB/DD-PTs on the one hand and the RTs on the other hand yielded numerous pronunciation variants in the resulting recognition lexica (though less than in the DD-PT lexica). The higher similarity between the original $[\text{CAN-PTs}]_d$, the $[\text{KB-PTs}]_d$ and the RTs led to fewer branches in the decision trees and fewer pronunciation variants in the resulting recognition lexica. As a consequence, the corresponding prior probabilities of the variants were intrinsically more robust than the probabilities in the data-driven lexica comprising more pronunciation variants per word.

Recall that we did not implement vowel reduction and deletion for the generation of the KB-PTs, and that we based our KB-PTs on canonical transcriptions without using supra-segmental information. We investigated whether the disregard of this knowledge in our knowledge-based transcription procedure made a substantial contribution to the discrepancies

between the KB-PTs (and consequently also the $[\text{KB-PTs}]_d$) and the RTs. This proved not to be the case; the missing vowel rules and the reformulation of the phonological processes did not hamper the pronunciation variation modelling in the knowledge-based transcriptions procedures to any substantial degree.

We obtained our best transcriptions by means of the procedure in which our fully canonical transcriptions were tuned towards the manually verified reference transcriptions by means of pronunciation variation modelling inspired by speech processes that were attested in the reference transcriptions. Apparently, learning intra-word and cross-word phonological processes from a small sample of real transcriptions works better than predicting the results of these processes from linguistic and phonetic knowledge. It remains to be explained why the KB-PTs were a less effective starting point for the learning process. We think that this is most likely due to a canonically-oriented bias in the RTs that was so strong that no other point of departure could close the gap. Thus, in order to approximate manually verified transcriptions resulting from the auditory verification of close-to-canonical example transcriptions (like in the Spoken Dutch Corpus), it is worthwhile learning the most obvious differences between the canonical and the reference transcriptions through the use of decision trees. One should bear in mind, though, that a canonical point of departure may be suboptimal to approximate RTs that are not based on a (similar) example transcription.

3.5.3 What about the remaining discrepancies?

The number of remaining discrepancies in the $[\text{CAN-PTs}]_d$ of the telephone dialogues (14.6% disagreement) and the read speech (8.1% disagreement) was only slightly higher than human inter-labeller disagreement scores reported in the literature. Recall that Binnenpoorte (2006) reported human inter-labeller disagreements between 14 and 11.4% on transcriptions of Dutch spontaneous conversations, and between 6.2 and 3.7% disagreements on transcriptions of Dutch read speech from the Spoken Dutch Corpus. In the context of the figures reported in Binnenpoorte (2006), a closer look at the 20 most frequent dissimilarities distinguishing our $[\text{CAN-PTs}]_d$ from the human RTs shows a comparable number of insertions and deletions, and a set of substitutions in which the mismatches between voiced and voiceless phones were dominant (see Table 3.5).

Similar disagreements were previously observed between different human transcribers who verified the same example transcription (Binnenpoorte et al., 2003). Therefore, we believe that our automatic transcription procedures have faced the same ‘mission impossible’ as humans when making broad phonetic transcriptions. The limited number of phonetic symbols available forces human transcribers and machines to classify auditory observations in a continuous space into discrete categories. For observations that are close to (hypothetical) category boundaries, forced choices inevitable cause a large proportion of disagreements. Fortunately, if for some application in which phonetic transcriptions must be used independent criteria can be formulated for classifying a fricative as voiced or unvoiced (to mention one of the most volatile phonetic differences in Dutch) it is probably quite easy to

train an acoustic classifier to re-label all fricatives in the corpus according to the new criteria. Most probably, such a re-labelling will be equally advantageous for manually verified broad phonetic transcriptions, for the same reason: they also involve classifications that may not fully adhere to the newly introduced criteria. Thus, we can conclude that we found a very quick, simple and cheap transcription procedure able to approximate manually verified phonetic transcriptions of a large speech corpus by training an automatic procedure on the basis of a relatively small set of data. Our procedure applied uniformly to well-prepared and spontaneous speech. It remains to be shown that the procedure is equally effective for manual transcriptions that are made in a way that is significantly different from the procedure used in the Spoken Dutch Corpus (and in most other large speech corpora, for that matter). However, the machine learning procedure on which our approach is based seems sufficiently general and powerful to approximate different types of transcriptions, as long as learning can be initialised from a starting point that is not too far from the eventual target.

Table 3.5: 20 most frequent mismatches between the $[CAN\text{-}PTs]_d$ and the RTs.

read speech						telephone dialogues					
substitutions		deletions		insertions		substitutions		deletions		insertions	
RT	$[CAN\text{-}PT]_d$	RT	$[CAN\text{-}PT]_d$	RT	$[CAN\text{-}PT]_d$	RT	$[CAN\text{-}PT]_d$	RT	$[CAN\text{-}PT]_d$	RT	$[CAN\text{-}PT]_d$
v	f	@	-	-	@	d	t	@	-	-	@
s	z	r	-	-	d	z	s	r	-	-	t
g	k	n	-	-	r	v	f	n	-	-	r
d	t	h	-	-	t	g	k	h	-	-	d
t	d			-	h	@	A			-	n
G	x			-	n	G	x			-	j
@	A					A	a				
z	s					t	d				
A	a					s	z				
@	a					f	v				

3.6 Conclusions

The aim of our study was to investigate whether existing automatic transcription procedures and combinations of such procedures can approximate the quality of manually verified phonetic transcriptions of speech. If such procedures would be able to do so, we would have a quick and cheap alternative to deploying human experts for the generation of the type of transcription of large speech corpora. We used ten automatic transcription procedures to generate a phonetic transcription of well-prepared speech (read-aloud texts) and of spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The resulting transcriptions were compared to the corresponding manually verified phonetic transcriptions from the Spoken Dutch Corpus.

Our results showed that, in order to approximate the quality of the manually verified phonetic transcriptions in the Spoken Dutch Corpus, one only needs an orthographic transcription, a canonical lexicon, a small sample of manually verified phonetic transcriptions, software for the implementation of decision trees and a standard continuous speech recogniser. Our study suggests that it is sufficient to verify the phonetic transcription of only a small portion of a corpus by hand in order to automatically generate similar transcriptions for the remainder of the corpus by means of decision trees. The best point of departure for such an automatic procedure will probably depend on the procedure by means of which the manual reference transcriptions were obtained.

CHAPTER 4

SEGMENT DELETION IN SPONTANEOUS SPEECH: A CORPUS STUDY USING MIXED EFFECTS MODELS WITH CROSSED RANDOM EFFECTS

Van Bael, C., Baayen, R.H., Strik, H. (to be submitted). Segment Deletion in Spontaneous Speech: A Corpus Study Using Mixed Effects Models with Crossed Random Effects.

Abstract

We studied the frequencies of phone and syllable deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the various linguistic and sociolinguistic factors represented in the transcriptions, word segmentations and metadata of the Spoken Dutch Corpus. In addition to providing insight into the frequencies of phone and syllable deletions and the factors influencing them, our study illustrates the new opportunities for analysing rich and therefore complex corpus data offered by a recently developed statistical modelling technique: the possibility to model the effects of random factors as crossed instead of nested with generalised linear mixed effects models.

We observed average phone and syllable deletion rates of 7.57% and 5.46% respectively. 20.32% of the words had at least one phone missing, and 6.89% of the words had at least one syllable deleted. The mixed effects models for phone and syllable deletion had several effects in common, which implies that both types of deletion are to a large extent influenced by the same factors. The strongest factors across both models were lexical stress, word duration and the segmental context of the syllable onset of the following word.

Keywords: *segment deletion, corpus linguistics, statistical modelling.*

4.1 Introduction

Over the years, large phonetically transcribed speech corpora have proven valuable resources for studying pronunciation variation. Switchboard (Godfrey et al., 1992; Greenberg et al., 1996) and the Buckeye Corpus of Conversational Speech (Pitt et al., 2005), to name just two examples, have proven useful for -among other things- creating an inventory of testified speech processes in everyday conversational English (Greenberg et al., 1996), studying the frequencies of these processes (Johnson, 2004) and investigating how these processes are influenced by various linguistic and sociolinguistic effects (e.g. Bell et al., 2003; Raymond et al., 2006). Historically, most phonetically transcribed speech corpora comprise (American) English. Therefore most corpus studies on pronunciation variation were conducted on English. The recent release of the 9-million-word Spoken Dutch Corpus (CGN; Oostdijk, 2002) now offers new opportunities for studying pronunciation variation in a language other than English, and for testing whether knowledge gleaned for American English also holds for another language. The CGN contains material from various speech styles, it was annotated with metadata including speaker characteristics, it was segmented at the word level and it was provided with an orthographic transcription, a broad phonetic transcription and POS tags. The word segmentations and phonetic transcriptions of a 1-million-word subset of the CGN, the so-called *core corpus*, were manually verified. We used the annotations, word segmentations

and metadata of the spontaneous telephone dialogues in this core corpus to study the deletion of phones and syllables in spontaneous Dutch.

The first aim of our study was to establish the frequencies of phone and syllable deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the linguistic and sociolinguistic factors reflected in the annotations, word segmentations and metadata of the CGN. Previous studies on (factors affecting) pronunciation variation were mostly conducted by means of controlled experiments. As a result they were often restricted with respect to the number and selection of investigated pronunciation processes, factors and words. Studies often considered one or a few speech processes at a time (e.g. Jurafsky et al. (2001) considered vowel reduction and final /t,d/-deletion and Raymond et al. (2006) considered word-internal /t,d/-deletion in American English). Likewise, studies often considered the effects of one or a few factors at a time (e.g. Bell et al. (2003) included the effects of preceding and succeeding disfluencies, word predictability, utterance position and gender) and/or a limited number or a selection of words (e.g. Bell et al. (2003) investigated the pronunciation of the 10 most frequent English function words and Keune et al. (2005) investigated the pronunciation of 14 Dutch words ending in '-lijk'). In our present study, we aimed at modelling the deletion of phones and syllables in spontaneous Dutch without any further restrictions on the nature of the deletions, factors and words under investigation: we modelled the effects of a wide range of linguistic and sociolinguistic factors on the phone and syllable deletions in the spontaneous telephone dialogues of the CGN.

An ancillary goal of our study was to explore the new opportunities for analysing rich and therefore complex corpus data offered by a recently developed statistical modelling technique: the possibility to model the effects of random factors as crossed instead of nested with generalised linear mixed effects models (Baayen et al., 2007). Mixed effects models allow for the inclusion of a mixture of fixed-effect factors and random-effect factors in the same model. Fixed-effect factors such as 'word class' are factors with a limited set of levels whose effects on a response variable (in our case: phone or syllable deletion) will remain unchanged in successive experiments. By contrast, random-effect factors such as 'speaker' or 'word' are randomly sampled from large populations. Because of this random selection, the effects of the different levels of random-effect factors (e.g. individual speakers or words) are usually not compared with each other but instead considered to describe the variation in effects of the population they were sampled from. Including fixed and random effects in one mixed effects model allows for generalising the effects of the fixed factors to the populations from which the random factors were sampled. In addition, including random effects alongside fixed effects allows for modelling any remaining speaker- or item-specific random effects that were not explicitly (as fixed effects) included in the model.

Mixed effects models, as argued in Pinheiro and Bates (2000), are robust with respect to missing data (this increases the stability of the estimates of fixed effects), they are easily applicable to designs with complex mixtures of factors, and at the same time they are parsimonious with regard to the number of parameters they need to estimate (random-effects factors only require one parameter: the variance from the group mean; fixed-effect factors require just as many parameters as there are levels). The possibilities to cope with missing

data, complex factorial designs, the ability to do all this in a computationally efficient way and to model both fixed-effects and random-effects factors in one model makes mixed effects models highly suitable for linguistic corpus studies. However, until recently, two or more random-effects factors in the same mixed-effects model could only be modelled as nested (e.g. the random-effects factor 'syllable' nested within the random-effects factor 'word' nested within the random-effects factor 'speaker'). This imposed serious limitations on the use of mixed-effects models for linguistic studies.

Nesting a factor A within a factor B implies the assumption that the levels of factor A are different within different levels of factor B. Thus, nesting a random-effects factor 'word' within a random-effects factor 'speaker' would force us to assume that the properties of individual words (e.g. word length, word class) differ so substantially for different speakers that their identity is no longer an issue across speakers. This assumption is demonstrably incorrect for many (psycho)linguistic data sets, and it could result in anti-conservative P-values for the fixed effects, making them too easily significant in addition to the random effects. In other words, this assumption could increase the risk of type I errors (i.e. erroneously considering an effect significant). Fortunately, it is now possible to model random effects in mixed-effects models as crossed instead of nested. This possibility allowed us to include random-effects factors such as 'speaker' and 'word' simultaneously as independent sources of random variation, and fixed-effects factors related to these random-effects factors (e.g. 'gender' for 'speaker', and 'word frequency' for 'word') as independent fixed effects. In turn, this enabled us to assess in one model which linguistic and sociolinguistic variables predict segment deletion over and above the random variation that came with the subjects and items (e.g. 'word') sampled.

This paper is organised as follows. Section 4.2 presents our data and methodology. In Section 4.3, we present and discuss the results of our analyses. Section 4.4 summarises our conclusions.

4.2 Methodology

4.2.1 Data preparation

We based our study on the annotations, word segmentations and metadata of the spontaneous telephone dialogues in the core corpus of the CGN. Excluding broken and (partially) unintelligible words, we obtained a dataset of 178,271 word tokens (8,539 types) with manually verified word boundaries, manually verified orthographic and broad phonetic transcriptions and POS tags. Similar to Johnson (2004), we generated a canonical representation of the material by concatenating the citation forms of the words. These citation forms (including syllable boundaries and lexical stress marks) were retrieved through lexicon lookup. The phones in the canonical representation and the manually verified broad phonetic transcription were aligned with ADAPT (Elffers et al., 2005). In this process, the syllable

boundaries and lexical stress marks of the canonical representation were copied onto the manually verified transcriptions of the CGN to identify phone and syllable deletions in these transcriptions (see Figure 4.1).

CT	Ei	G @	l @ k
PT	Ei	- -	l @ k

Figure 4.1: Identification of two phone deletions and one syllable deletion through the alignment of a canonical transcription (CT) and a broad phonetic transcription (PT).

For every canonical syllable, we retrieved linguistic information at the utterance[1], word and syllable level from the orthographic and canonical transcriptions, the POS tags and the word segmentations. For every canonical phone, we retrieved the same information and additional information at the phone level. Sociolinguistic (speaker) information was extracted from the metadata. All linguistic and sociolinguistic information (e.g. the current word was an adjective and spoken by a male subject) was stored in a separate *information vector* for every canonical phone and syllable to allow for modelling the effects on phone and syllable deletion of the various linguistic and sociolinguistic factors (in the above example the factor 'word class' with the level 'adjective' and the factor 'gender' with the level 'male').

At the utterance level, we considered the duration (in ms excluding long silent pauses) and the number of canonical phones and syllables. From this information we computed the articulation rate in phones and syllables per second. At the word level, we considered the word identity, the word duration (in ms excluding silent pauses), the number of canonical phones and syllables, the position in the utterance (initial, final, initial-final, mid), the word class (nouns, verbs, adjectives, adverbs, pronouns, interjections, articles, numerals, conjunctions, prepositions), and the number of times the word was previously uttered by one of the interlocutors and by the current speaker (to model the effects of given/new information). We also considered the word's frequency and the mutual information of the word and its neighbours (both computed on the orthographic transcription of the 544,215 word tokens in the telephone dialogues of the CGN that were not included in the core corpus), whether the word preceded a long silence (>250 ms) or a disfluency (repetition, broken word, filled pause) and whether the following word started with a consonant or a vowel. At the syllable level, we considered the syllable identity, the syllable's position in the word (initial, final, initial-final, mid), the number of canonical phones, and whether the syllable had lexical stress (retrieved through lexicon lookup). At the phone level, we considered the phone identity, its position in the word (initial, final, initial-final, mid), syllable (initial, final, initial-final, mid) and in the consonant/vowel structure of the syllable (e.g. CC_V), whether the phone was part of the syllable's onset, nucleus or coda and whether the phone had lexical stress (retrieved through

[1] Utterances were defined as stretches of speech that were marked with capitals and punctuation marks in the orthographic transcription of the CGN.

lexicon lookup). In addition, we considered the identity of the speaker, his or her gender, age (year of birth), regional background (the region the speaker spent most of the time between the age of 4 and 16) and level of education (high, mid, low). In the last field of each information vector, we marked whether the phone or syllable was deleted. Like Johnson (2004) in his study on phone and syllable reductions in conversational English, we considered syllables deleted if their syllabic nucleus was absent. Contrary to English, Dutch normally does not have syllabic nasals, laterals or rhotics. Therefore we considered syllables deleted if a vocalic nucleus was no longer present.

4.2.2 Analyses

We first counted the number of phone and syllable deletions to gain insight into the frequencies of such deletions in spontaneous Dutch. The results of these quantitative analyses are presented in Section 4.3.1. Subsequently, we fitted two generalised linear mixed effects models with a logistic link function to the information vectors: a model for phone deletion and a model for syllable deletion. We assumed binomial variance. Both models were defined by sequentially including every linguistic and sociolinguistic factor from the information vectors in the model. A factor was only retained if it contributed significantly ($p < .05$) to the model's goodness of fit. Factors were pruned from the model if their contribution was no longer significant after the inclusion of additional factors. Goodness of fit was assessed with Somers' Dxy, a rank correlation between predicted probabilities and observed responses which is closely related to the receiver operating characteristic curve area (Harrell et al, 1996). The results of the statistical analyses are presented in Section 4.3.2. All statistical computations were conducted with the `lme4` package (Bates, 2005; Bates and Sarkar, 2005) for R (R development core team, 2005). In order to keep building the models computationally feasible, we fitted models on a randomly selected 10% subset of the material.

Our study required the inclusion of fixed-effect and random-effect factors. Recall that fixed-effect factors such as 'word class' are factors whose levels (e.g. 'adjective') are selected from a limited set of values, and repeatable in successive experiments (the word 'beautiful' will always be an adjective). By contrast, random-effect factors such as 'speaker' are factors whose levels (in this case: individual speakers) are randomly sampled from a larger population. In statistical modelling, especially when fitting models with a large number of fixed-effect factors, fixed-effect factors with a large number of levels (e.g. the factor phone, which had 42 levels to account for the different phones in the transcriptions) are often treated as random-effect factors. This drastically reduces the number of estimated parameters in the model: fixed-effect factors require as many parameter estimates as there are levels (thus 42 for the factor 'phone') whereas random variables can be modelled by one parameter: the variance from the group mean. Decreasing the number of parameter estimates in a model increases the transparency and hence the interpretability of the model. Post-hoc inspection of the best linear unbiased predictors for the random-effect factors makes it possible to investigate which levels of each random-effect factor (e.g. the level /@/ for the factor 'phone') influenced the deletion

of phones and syllables more or less than the random intercept representing the mean effect of all the levels of the factor (Baayen, forthcoming).

In our study, 'speaker', 'word' and 'syllable' were straightforward cases of random-effect factors because they were all randomly sampled from large populations. In addition, in order to limit the number of parameters in the models, we treated the otherwise fixed-effects factors 'regional background of the speaker' (16 levels), 'phone' (42 levels) and 'the position of the phone in the consonant/vowel structure of the syllable' (e.g. CC_V, 57 levels) as random effects. All remaining factors in the enumeration of factors below were obvious fixed-effect factors and treated as such.

4.3 Results

4.3.1 Frequencies of segment deletions

Phone deletions

We counted 42,556 phone deletions out of 562,294 phones in 85,050 content and 93,221 function words[2]. This implies an overall phone deletion rate of 7.57%. Table 4.1 lists the deletion rates of the individual phones (with 95% confidence intervals), and the proportion of all phone deletions accounted for by the deletion rates of the individual phones. 83.69% of all phone deletions concerned deletions of one of the 7 following phones: /@/ (22.91% of all deletions), /r/ (19.59%), /n/ (13.12%), /t/ (12.27%), /l/ (5.74%), /h/ (5.39%) and /d/ (4.66%). This was not just because these phones are common in Dutch; the individual deletion rates of these phones proved higher than the deletion rates of other phones. We found the following proportion of deletions for each of these phones: /r/ (28.79% - 95% confidence interval: 1.05), /h/ (21.25% - 1.55), /@/ (16.10% - 0.59), /n/ (12.40% - 0.61), /l/ (12.29% - 0.92), /t/ (11.40% - 0.58), /d/ (7.06% - 0.6).

Looking at the data from a more general perspective, the deletion of plosives (/p/, /b/, /t/, /d/, /k/, /g/) and sonorants (/m/,/n/,/N/, /l/, /r/, /j/, /w/) took up a large portion of all phone deletions: 18.06% resp. 42.09% (60.15% in all). This is in line with Johnson (2004), who also reported most phone deletions in his sample of American English to concern plosives and sonorants. The deletion of full vowels (including diphthongs, nasalised and loan vowels) accounted for only 6.73% of all phone deletions. The deletion of /@/ however in itself accounted for 22.91% of all phone deletions. Mind that, since we considered the deletion of citation form phones, the low average deletion rate of full vowels (6.73%) and the high average deletion rate of /@/ (22.91%) do not imply that the deletion of full vowels proceeded through a two-step procedure via reduction to schwa and subsequent deletion of schwa. In other words, the lower figure for vowel deletions (6.73%) was not because vowels were first

[2] Nouns, verbs, adverbs and adjectives were considered content words. All other words were treated as function words.

reduced to /@/ and afterwards deleted and counted as deletions of /@/. Rather, the fact that 22.91% of all phone deletions was accounted for by deletions of /@/ simply implies that /@/, which often occurs in syllables without lexical stress and without prosodic accent, is highly susceptible to deletion. To complete this description of average phone deletion rates, the deletion of fricatives (/f/, /v/, /s/, /z/, /x/, /G/ and /h/) accounted for the remaining 10.22% of the phone deletions.

Table 4.1: Deletion rates of phones (%del, with 95% confidence intervals) and proportions of phone deletions accounted for by the individual phone deletion rates (% of all deletions). SAMPA symbols are used. Classes: PLosives, FRicatives, SOnorants, Short Vowels, Long Vowels, DIphthongs, LoanVowels, Nasalised Vowels, @.

class	phone	# tokens	% del	% of all deletions	class	phone	# tokens	% del	% of all deletions
PL	p	7,213	1.14 (0.51)	0.19	SV	I	15,387	5.68 (0.74)	2.05
	b	6,520	1.12 (0.53)	0.17		E	21,014	1.95 (0.38)	0.96
	t	45,814	11.40 (0.58)	12.27		A	26,026	0.97 (0.24)	0.59
	d	28,074	7.06 (0.60)	4.66		O	10,827	1.43 (0.46)	0.36
	k	22,128	1.45 (0.32)	0.75		Y	3,834	2.58 (1.03)	0.23
	g	68	5.88 (13.23)	0.01	LV	i	11,869	1.70 (0.47)	0.47
FR	f	4,346	2.37 (0.93)	0.24		y	1,909	1.00 (0.97)	0.04
	v	9,524	0.68 (0.34)	0.15		e:	15,479	0.81 (0.29)	0.30
	s	19,920	5.19 (0.62)	2.43		2	1,002	0.00	0.00
	z	9,998	0.43 (0.27)	0.10		a:	27,407	1.59 (0.30)	1.02
	S	253	0.00	0.00		o:	14,589	0.99 (0.33)	0.34
	Z	132	0.00	0.00		u	5,958	0.55 (0.40)	0.08
	x	16,507	3.76 (0.59)	1.46	DI	Ei	7,213	0.78 (0.42)	0.13
	G	3,551	5.32 (1.51)	0.44		9y	1,470	0.14 (0.53)	0.00
	h	10,802	21.25 (1.55)	5.39		Au	3,719	0.48 (0.48)	0.04
SO	N	2,724	0.51 (0.59)	0.03	LoV	E:	29	0.00	0.00
	m	17,607	1.89 (0.41)	0.78		9:	6	0.00	0.00
	n	45,055	12.40 (0.61)	13.12		O:	7	0.00	0.00
	J	26	0.00	0.00	NaV	E~	2	0.00	0.00
	l	19,882	12.29 (0.92)	5.74		A~	71	50.70 (23.99)	0.08
	r	28,950	28.79 (1.05)	19.59		O~	4	0.00	0.00
	w	15,181	5.91 (0.76)	2.11		Y~	-	-	-
	j	19,636	1.54 (0.35)	0.71	@	@	60,561	16.10 (0.59)	22.91

When assessing phone deletion at the word level, we found that 17.56% of the words had one phone missing, 2.16% had two phones missing, and 0.60% had three or more phones missing. We found relatively more phone deletions in function words (18,382 out of 213,975 phones - 8.59%) than in content words (24,174 out of 348,319 phones - 6.94%), but at the same time we observed more individual content words than function words with phone deletions: 22.76% of the content words and 18.10% of the function words had at least one phone missing. This implies that the phone deletions in the function words were concentrated in a proportionally smaller subset of the words than was the case with the content words. This fits in with Johnson's observation for English that the function words in the Buckeye Corpus

deviate more from their citation forms (in terms of phone substitutions, deletions and insertions) than the content words, and with previous observations that function words, as opposed to content words, are more susceptible to segment reduction and deletion. As Raymond et al. (2006:62) summarises, function words are generally less salient than content words because they have lower semantic content, they are more common, they are generally shorter and they usually occur in unaccented positions. As a result, function words are often reduced and/or cliticised in connected speech. Content words on the other hand are often multisyllabic and more likely to receive prosodic accent, which results in fuller pronunciation variants.

In general, the phone deletion rates we found for Dutch were a little lower than the figures reported for American English in the Buckeye Corpus. Johnson (2004) reported that a little over 20% of the words in his dataset had one phone deleted (we found 17.56%) and that 5% of the words had two or more phones missing (we found 2.66%). However, just like Johnson (2004), we have to conclude that, given the short average word length (in our study 3.14 canonical phones per word – the top 50 of most frequent word types, which accounted for 56.47% of all word tokens, contained one word type with 4 canonical phones and 49 types with three or less phones), the proportion of words with one or more missing phones is remarkably large. Table 4.2 illustrates how the phone deletions were distributed in words of different length (in terms of the number of syllables and phones per word).

Table 4.2: Frequencies of phone deletion in words of different length (95% confidence intervals between brackets).

	content words		function words	
# syllables/word	# phones	% phone deletion	# phones	% phone deletion
1	155,463	6.82 (0.25)	191,208	8.66 (0.25)
2	108,329	6.02 (0.28)	18,813	8.39 (0.80)
3	61,444	9.27 (0.46)	2,113	5.35 (1.97)
4	17,515	5.85 (0.70)	1,351	6.88 (2.78)
5 or more	5,568	5.85 (1.25)	490	6.73 (4.67)
# phones/word	# phones	% phone deletion	# phones	% phone deletion
1	24	0.00	10,080	6.14 (0.95)
2	23,520	3.21 (0.46)	111,294	7.24 (0.31)
3	101,862	7.74 (0.33)	66,978	11.56 (0.49)
4	52,624	6.13 (0.41)	11,492	5.89 (0.87)
5	46,035	5.78 (0.43)	5,220	4.96 (1.20)
6	42,222	8.20 (0.53)	4,284	8.52 (1.70)
7	27,377	7.23 (0.62)	2,191	22.23 (3.53)
8	22,664	8.81 (0.74)	280	7.86 (6.72)
9	13,860	7.61 (0.89)	180	1.67 (4.76)
10 or more	18,131	6.38 (0.72)	1,976	7.79 (2.42)
total	348,319	6.94 (0.17)	213,975	8.59 (0.24)

Table 4.2 shows both for content and function words a steady decrease in the phone deletion rates in monosyllabic and bisyllabic words and words with 4 syllables or more. Function words, as already indicated, in general showed higher phone deletion rates than content words. This did not hold for trisyllabic words though: trisyllabic content words were remarkably more susceptible to phone deletion than trisyllabic function words. Moreover, trisyllabic content words were remarkably more susceptible to phone deletion than other content words, whereas trisyllabic function words had relatively fewer phones deleted than the other function words. Closer inspection of the deletions in trisyllabic words per word class showed that in particular trisyllabic (and often very common) adverbs such as 'allemaal', 'helemaal', 'inderdaad' and adjectives such as 'natuurlijk', 'eigenlijk', 'waarschijnlijk', 'allerlei' were highly susceptible to phone deletion (see Table 4.3 for the frequencies of phone deletions in trisyllabic words per word class).

The last column in Table 4.3 summarises the phone deletion rates computed over all phones per word class. These figures show that in particular articles, pronouns and conjunctions were prone to deletion. This is in line with the findings of Greenberg (1998) for American English: these word classes were found to depart most regularly from their canonical form. We found that interjections, nouns and numerals were least prone to phone deletion. The low deletion rate of the interjections can be largely explained by the frequent use of filled pauses, which by default were transcribed by means of their citation form /@/. The low deletion rate of both nouns and numerals can be explained by the high information valence associated with these words. Numerals are probably the class of function words with the highest information value, which explains why they were less prone to phone deletions, as were the nouns.

Table 4.3: Frequencies of phone deletion in trisyllabic words (fourth column) and all words (last column) (95% confidence intervals between brackets).

	word class	# phones in trisyllabic words	% phone dels in trisyllabic words	# phones in all words	% phone dels in all words
content	noun	19,930	5.05 (0.61)	83,704	3.98 (0.27)
	verb	14,727	5.39 (0.74)	116,702	6.76 (0.29)
	adjective	15,272	11.63 (1.02)	58,544	7.72 (0.43)
	adverb	11,515	18.41 (1.42)	89,369	9.43 (0.38)
	total	61,444	9.27 (0.46)	348,319	6.94 (0.17)
function	pronoun	976	2.56 (2.12)	81,984	10.82 (0.43)
	interjection	164	7.32 (8.71)	45,613	2.89 (0.31)
	numeral	589	8.32 (4.65)	8,379	4.66 (0.91)
	conjunction	119	11.77 (12.46)	31,935	11.31 (0.70)
	preposition	265	4.91 (5.69)	32,405	7.76 (0.59)
	article	-	-	13,659	12.69 (1.12)
	total	2,113	5.35 (1.97)	213,795	8.59 (0.24)

The figures in Table 4.2 also show that content words with two phones or less were less susceptible to phone deletion than content words with three or more phones. This can probably to a large extent be explained by the frequent use of discourse words like 'ja' (yes) and 'mm' (uhu), which often constitute a speech utterance of their own and are therefore well pronounced. The deletion rates of content words with three or more phones were relatively stable. Johnson (2004) observed a similar effect for American English: 1, 2, and 3-phone content words showed increasing deletion rates, content words with more phones showed relatively stable deletion rates between 7% and 12%. Johnson (2004) reported a similar tendency for English function words: he found a constant rise in deletion rates for 1, 2 and 3-phone function words, and a relatively stable deletion rate for words with more phones. We did also notice increasing deletion rates for 1, 2, and 3-phone function words, with a remarkably high number of phone deletions in 3-phone function words (11.56%). This high deletion rate was largely due to the frequent occurrence of final /r/-deletions in common words such as 'naar' (towards), 'daar' (there), 'maar' (but), 'hoor' (discourse word) and 'voor' (for) and due to the frequent use of words such as 'm'n' (from: 'mijn' - my), 'z'n' (from 'zijn' - his) and 'd'r' (from 'haar' – her, or from 'daar' - there) in which the canonical vowel nucleus (i.c. /@/) was deleted. These observations largely explain the high average phone deletion rates of /r/ and /@/ reported in Table 4.1, and the high average syllable deletion rates in monosyllabic function words that will be reported below.

Contrary to the observations in Johnson (2004), the phone deletion rates in longer function words (more than three phones) were all but stable. In particular the average phone deletion rate in 7-phone function words and (to a lesser extent because of the limited number of phones this figure was based upon) the average phone deletion rate in 9-phone function words were clearly distinct from the deletion rates of the remaining function words. The 22.23% phone deletion rate in 7-phone function words is largely due to heavy reductions in the preposition 'volgens', which occurred 170 times (and accounted for 54.31% of the 313 7-phone function word tokens) and which had an average phone deletion rate of 34.96%. This large deletion rate can be explained by the fact that in 155 out of 170 cases, the preposition 'volgens' preceded the pronoun 'mij' in the multi-word expression 'volgens mij' (according to me). Previous research on multi-word expressions has shown that (function) words in such frequent N-grams often undergo heavier reduction than the same words in other contexts (Bell et al., 2003; Binnenpoorte et al., 2005).

Syllable deletions

We observed that 12,534 out of 229,670 syllables (5.46%) were deleted. As with the frequencies for phone deletion, we observed relatively more syllable deletions in function words (6,592 out of 98,690 syllables - 6.68%) than in content words (5,964 out of 130,981 syllables - 4.54%). 7.09% of the function words had 1 syllable missing, and a negligible 0.01% had 2 syllables missing. 6.33% of all content words was pronounced with 1 syllable missing, 0.33% had 2 or more syllables missing. These results are not in line with the

observations for American English reported in Johnson (2004), where relatively more content words (6%) than function words (4.5%) were reported to have at least one missing syllable. The relatively high syllable deletion rate in function words in our data can be explained by the frequent use in Dutch of clitics such as ‘’k’ (from: ‘ik’ - I), ‘d’r’ (from ‘haar’ – her, or from ‘daar’ - there), ‘m’n’ (from: ‘mijn’ - my), ‘’t’ (from: ‘het’ – it) and ‘’n’ (from ‘een’, - a). The citation forms for these shortened word forms contain a /@/ as their syllabic nucleus, but this /@/ was often deleted in our material. We counted 6,963 such /@/ deletions in function word clitics. This accounts for the deletion of 2.62% of the syllables in the function words. The remaining 4.06% (6.68% – 2.62%) syllable deletion rate is comparable to the 4.5% syllable deletion rate reported for function words in American English, and below the 4.54% phone deletion rate we counted in the content words[3]. Ignoring the deletion of citation form schwas due to cliticisation, we found fewer syllable deletions in function words (4.06%) than in content words (4.46%).

Table 4.4: Frequencies of syllable deletion (95% confidence intervals between brackets).

content words					
# canonical syl	totals	no syl del (%)	-1 syl del (%)	-2 syl del (%)	-3 syl del (%)
1	52,638	98.35 (0.22)	1.65 (0.22)		
2	43,460	94.34 (0.44)	5.66 (0.44)		
3	25,224	91.80 (0.68)	6.42 (0.61)	1.78 (0.33)	
4	7,324	94.51 (1.06)	4.61 (0.98)	0.82 (0.43)	0.04 (0.12)
5	1,795	93.70 (2.31)	4.35 (1.95)	1.78 (1.29)	0.17 (0.49)
6	360	95.00 (4.85)	3.89 (4.37)	1.11 (2.66)	
7	105	98.10 (7.05)	1.90 (7.05)		
8	56	98.21 (10.72)	1.79 (10.72)		
9	18	100			
function words					
# canonical syl	totals	no syl del (%)	-1 syl del (%)	-2 syl del (%)	-3 syl del (%)
1	88,427	93.02 (0.34)	6.98 (0.34)		
2	8,632	95.69 (0.87)	4.19 (0.86)	0.12 (0.16)	
3	963	96.47 (2.46)	3.12 (2.33)	0.42 (1.01)	
4	484	97.52 (3.06)	2.48 (3.06)		
5	160	96.88 (6.37)	3.13 (6.37)		
6	24	100			

Table 4.4 shows the distribution of syllable deletions over N-syllable content and function words. Not surprisingly, the deletion of just one syllable was more common than the deletion of more syllables. Compared to the corresponding deletion rates reported by Johnson (2004) for conversational English, however, we found much higher deletion rates for Dutch

[3] We also counted 487 cliticised content words (mostly adverbs such as ‘’ns’ (from: ‘eens’ - once) and ‘d’rop’ (from: ‘daarop’ - on there)). However, /@/-deletion in these words only accounted for the deletion of 0.08% of the syllables in the content words. The impact on the overall syllable deletion rate in the content words was therefore not at all comparable to the impact of the cliticised function words on the syllable deletion rate in function words.

monosyllabic function words, and much lower deletion rates for Dutch content words with three syllables or more and function words with two syllables or more. Hardly any of the syllabic nuclei in the monosyllabic content words were deleted (1.65%). The most frequent syllabic nucleus deleted in monosyllabic content words was the [I]. The phone was deleted 850 times, of which 844 times in the singular verb form 'is' of 'zijn' (to be). It accounted for 39.01% of the deleted nuclei in monosyllabic content words. Contrary to the resistance to deletion of vocalic nuclei in monosyllabic content words, we found that 6.98% of the syllabic nuclei in monosyllable function words were deleted. /@/ was the most frequently deleted syllabic nucleus in monosyllabic function words. The deletions of schwas accounted for 77.26% of all deleted nuclei in monosyllabic function words. 55.43% of all deleted schwas in these words occurred in clitics. The second most deleted syllabic nucleus in monosyllabic function words was [I] in the frequently used personal pronoun 'ik' (I) and the preposition 'in' (in). It accounted for 6.88 % of all deleted nuclei.

4.3.2 Modelling segment deletion

Phone deletions

We fitted a generalised linear mixed effects model to the data, in the way as described in Section 4.2.2, with speaker, phone, syllable, word, syllabic structure and regional background of the speaker as crossed random effects and phone deletion as response variable. Somers' Dxy of the final model was equal to 0.86, which corresponds to a receiver operating characteristic curve area of 0.93 and which indicates that the model provided a good fit to the data. Most fixed-effects predictors were significant at at least the 0.01 level. Inclusion of the random effects in the model was supported by likelihood ratio tests (ANOVA tests, all p-values < 0.05). In addition to the effects of the phone identity ($\hat{\sigma}$= estimated standard deviation of the random effect = 1.56), syllabic structure ($\hat{\sigma}$ = 1.06) and the regional background of the speaker ($\hat{\sigma}$ = 0.16) which were all treated as random-effects factors to limit the number of parameter estimates in the model (see Section 4.2.2) we observed main effects for eight fixed-effects factors over and above the random variation that came with the speakers ($\hat{\sigma}$ = 0.24) and the items (words ($\hat{\sigma}$ = 0.86) and syllables ($\hat{\sigma}$ = 0.93)) sampled.

We found a large main effect of word class. This is not surprising considering the large differences between the average phone deletion rates for the word classes presented in the last column of Table 4.3. We also found a large main effect of lexical stress: syllables with lexical stress were less likely to undergo phone deletion (Z = -10.99, p < 0.001). This effect has previously been described in the literature (see Raymond et al. (2006:65) and the references therein). Similar to previous studies such as Bell et al. (2005), we observed a main effect of word frequency: more frequent words were more susceptible to phone deletion (Z = 4.09, p < 0.001). In addition, we observed a main effect for the position of the phone in the syllable. We found that phones were much less likely to be deleted in onset (Z = -6.22, p < 0.001) and

nucleus positions (Z = -4.99, $p < 0.001$) than in coda positions. This is in line with findings for American English: Greenberg (1998) observed that syllable onsets are generally similar to their canonical form and thus usually preserved, that syllabic nuclei are usually preserved because they constitute the cores of syllables, and that phones in coda positions are often deleted or at least substituted with surface realisations that are homorganic with the phone of the following syllable's onset. Since in our study, we did not model cross-word phonological processes in the canonical representation of the material, it may very well be that we found many phone deletions in coda position as a result of subsequent assimilation and degemination with the onset of the following word. This probably also explains the additional main effects of the position of the phone in the word (initial phone, last phone, somewhere in between), and of the segmental context of the following word (first phone consonant or vowel). We found that deletions further into the word were more likely (with comparable deletion likelihoods for phones in word-final position (Z = 8.38, $p < 0.001$) and all remaining non-initial positions (Z = 10.33, $p < 0.001$)), and we found that words starting with a vowel were less likely to induce phone deletion in the preceding word than words starting with a consonant (Z = -3.63, $p < 0.001$). In addition to these main effects, we observed that utterances with more canonical phones were less likely to undergo phone deletion (Z = -3.09, $p < 0.01$). Remarkably in the context of this effect but not in contrast with it is that words with more canonical phones were more likely to undergo phone deletion (Z = 17.59, $p < 0.001$). At last, we noticed a large effect of the (log of the) duration of the word (in ms). The longer the actual duration of a word, the smaller the chance that phones were deleted (Z = -46.71, $p < 0.001$).

Inspection of the best linear unbiased predictors (BLUPs in short) for the phone random effect (i.e. the by-phone adjustments to the overall intercept) revealed that, according to our model, /h/, /d/ and /@/ were most likely to be deleted, and that /p/, /k/ and /N/ were least easily deleted. The ordering of the BLUPs from highest likelihood for phone deletion to lowest likelihood for deletion generally agreed with the ordering of the individual deletion rates of the phones in Table 4.1. Inspection of the BLUPs for syllabic structure showed that, according to the model, phones were most likely to be deleted in C_CV, _C, CVC_C, and CVCC_ structures and most resistant to deletion in CV_C, CCV_C and V_ structures. Whereas we didn't find support for the high deletion rates of phones in C_CV position, most of these findings were confirmed by our frequency counts. The susceptibility of the vocalic nucleus to deletion in VC syllables, which is clearly deviant from the findings of Greenberg (1998) for American English, can to a large extent be explained by the frequent use of clitics in Dutch. The deletion of schwa in such clitics accounted for 58.44% of all deletions in VC (or: _C) clusters. The deletion of schwa in what was in the orthography transcribed as the full form of the indefinite article 'een' (/@n/ - a) accounted for another 7.58%, and the deletion of /I/ in the singular tense 'is' (/Is/ - from: zijn - to be) and the personal pronoun ik (/Ik/ - I) accounted for another 22.25% of the deletions in VC clusters.

Syllable deletions

The generalised linear mixed effects model for syllable deletion was fitted with speaker ($\hat{\sigma}$ = 0.29), word ($\hat{\sigma}$ = 1.73) and syllable ($\hat{\sigma}$ = 2.00) as random-effects factors and syllable deletion as response variable. We considered including the regional background of the speaker and the syllabic structure as additional random effects, but likelihood ratio tests proved that these factors had no significant additional effect on syllable deletion. Because we were working at the syllable level (i.e. with the syllable as basic unit) including the random-effects factor phone was not an issue. Somers' Dxy was equal to 0.92 and the receiver operating characteristic curve area was equal to 0.96. These results indicate that also the syllable model fitted the data very well. All fixed-effects predictors were significant at at least the 1% level. Similar to our model for phone deletion, we observed main effects of word class and lexical stress. As expected, syllabic stress rendered syllable deletion less likely (Z = -12.22, p < 0.001). As opposed to what we saw in the model for phone deletion, words starting with a vowel increased the likelihood of syllable deletion in the previous word (Z = 12.45, p < 0.001). Considering the high deletion rate of schwas in our material, it is not unlikely that many schwas in unstressed and unaccented word-final syllables were deleted to ease the articulatory transition to the vowel of the next word. We also found a main effect of the (log of the) number of canonical syllables in the utterances: as the number of syllables increased, the likelihood of syllable deletion increased as well (Z = 2.90, p < 0.01). Somewhat related, we noticed that a higher articulation rate (expressed in canonical syllables per second) rendered syllable deletions more likely (Z = 4.27, p < 0.001). Furthermore, we found that syllables in utterance-initial words were more prone to deletion (Z = 4.04, p < .0.001) and that, similar to what we found for phone deletion, longer word durations are strong cues for the preservation of syllabic nuclei (Z = -32.49, p < 0.001).

Discussion

Fitting mixed-effects models for phone and syllable deletion taught us which linguistic and sociolinguistic factors in the annotations, word segmentations and metadata of the CGN affect phone and syllable deletion over and above the random variation that came with the speakers, words and syllables we sampled. Because we were able to include random-effects factors such as speaker, word and syllable identity as crossed instead of nested (the same indirectly also holds for all the other factors related to these random-effects factors), we were able to investigate in a methodologically sound way the relative effect of every factor over and above the effects of the other factors listed in Section 4.2. This makes analyses of mixed-effects models with crossed random effects interesting for two reasons. In our study, it was interesting to analyse which factors were significant in the models, but it was equally interesting to see that the (potential) effects of factors which were previously reported to influence segment deletion were ‘covered’ by other factors. For example, mutual

information (word predictability) which was previously reported to influence phone deletions (e.g. Bell et al., 2005) did not appear in our final model definitions, and word frequency was only significant in the phone deletion model. This may be due to several reasons. For example, we computed word frequency and mutual information on 'only' 544,215 words, and we chose to keep our models easily computable by not including correlations between factors. Computing estimates for word frequency and mutual information on a larger dataset and including correlations in the models may eventually render factors like word frequency and mutual information significant. We leave these issues open for further research. In any case, the absence of effects of e.g. word frequency (only in the syllable deletion model) and mutual information (in both model definitions) does not mean that these factors do not affect phone and syllable deletion. Rather, this implies that in our model definitions other factors showed a stronger effect on the deletion of phones and syllables. Actually, word frequency was part of the syllable model definition until we included 'word identity' as random effects factor. In both models, the effects of mutual information were probably covered by word frequency. Such knowledge is unlikely to be gained in controlled experiments on selected data sets aimed at studying the effects of one or a few factors at a time, but it can be of interest for pronunciation variation modelling of everyday conversational speech.

A comparison of the models for phone and syllable deletion shows that both types of segment deletion are to a large extent influenced by the same factors. We noticed main effects of word class, lexical stress, the segmental context of the following word (starts with a consonant or vowel), and the duration (in ms) of the pronunciation of the word. An interesting difference between the two models was the effect of the speaker's regional background (in addition to the by-speaker adjustments) on phone deletion. We did not notice such an effect on syllable deletion. This may indicate that more substantial deletions as gauged by syllable deletion (e.g. /Eix@l@k/ ('eigenlijk') reducing to /Eik/) are common to all speakers (but see Keune et al. (2005) for the presence of sociolinguistic variation for a subset of words in the CGN ending in the suffix -lijk), irrespective of their regional background.

4.4 Conclusions

We studied the frequencies of phone and syllable deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the interplay of linguistic and sociolinguistic factors than can be retrieved from the word segmentations, annotations and metadata in large annotated speech corpora such as the Spoken Dutch Corpus. We modelled the effects of various factors by means of a recently developed statistical modelling technique: the possibility to model the effects of random factors as crossed instead of nested with generalised linear mixed effects models.

We found average phone and syllable deletion rates of 7.57% and 5.46% respectively. 22.76% of the content words and 18.10% of the function words had at least one phone missing, and 6.66% of the content words and 7.10% of the function words had at least one

syllable missing. The high syllable deletion rate in function words was largely due to the frequent use of clitics in Dutch. Even though our figures of phone and syllable deletion are lower than the figures reported in Johnson (2004) for American English, our analyses also suggest that phone and syllable deletions are common in everyday conversational Dutch.

The mixed effects models for phone and syllable deletion had several effects in common, which implies that both types of deletion are to a large extent influenced by the same factors. An interesting difference between the models was that phone deletion appeared influenced by the regional background of the speaker (in addition to the by-speaker adjustments included in both the phone and syllable models) whereas syllable deletion was not. The strongest factors across both models were lexical stress, word duration and whether the following word starts with a vowel or a consonant.

Our study illustrates new opportunities for analysing rich and complex corpus data by means of generalised linear mixed effects models with crossed random effects. The use of such statistical models is useful for exploratory research like ours (we investigated the effects of a wide variety of linguistic and sociolinguistic factors on segment deletion) and for hypothesis testing (e.g. for testing whether a specific factor has an effect on a specific speech process in addition to the effects of other factors). Generalised linear mixed effects models in general are useful for studying complex corpus data because they can cope with missing data, because they can model the effects of many factors in one model, and because they are parsimonious with regard to the number of parameters they need to estimate for their random and fixed-effects factors. The recent possibility to include random-effects factors as crossed instead of nested now makes it possible to include several random-effects simultaneously as independent sources of random variation. This no longer presupposes the assumption that the levels of random-effects factors are dependent of each other, and this makes it possible to assess in a methodologically sound way in one model which linguistic and sociolinguistic variables predict segment deletion over and above the random variation that comes with the sampled levels of the random effects. Including random effects as crossed instead of nested decreases the risk of type I errors because it prevents fixed-effects factors of modelling random variation that should be better modelled by means of by-subject and by-item adjustments (i.e. by means of random-effects factors). Through this new modelling technique, linguistic phenomena such as segment deletion can now be studied in a methodologically sound way in corpus data as a function of the interplay of many factors instead of in controlled experimental environments designed for studying the effects of one or a few factors at a time.

CHAPTER 5

SPEAKER CLASSIFICATION BY MEANS OF ORTHOGRAPHIC AND BROAD PHONETIC TRANSCRIPTIONS OF SPEECH

Reformatted from:

Van Bael, C., Halteren, H. van (in press). Speaker Classification by means of Orthographic and Broad Phonetic Transcriptions of Speech. In: Müller, C., Schötz, S. (Eds.) *Speaker Classification*. *Lecture Notes in Computer Science/Artificial Intelligence*, Vol. 4343, Springer, Heidelberg – Berlin - New York.

Abstract

We investigated whether a classification algorithm originally designed for authorship verification can be used to classify speakers according to their gender, age, regional background and level of education by investigating the lexical content and the pronunciation of their speech. Contrary to other speaker classification techniques, our algorithm did not base its decisions on direct measurements of the speech signal; rather it learned characteristic speech features of speaker classes by analysing the orthographic and broad phonetic transcriptions of speech from members of these classes. The resulting class profiles were subsequently used to verify whether unknown speakers belonged to these classes.

Keywords: *Speaker Classification, Linguistic Profiling, Orthographic Transcriptions, Broad Phonetic Transcriptions.*

5.1 Introduction

Human listeners can rely on multiple modalities to determine a speaker's gender, age, regional background and -be it with less confidence- his or her level of education. Visual as well as auditory input can provide us with cues about a speaker's gender and age. In addition, auditory input can teach us a great deal about a speaker's regional background and level of education.

The aim of our study was to investigate whether Linguistic Profiling, a supervised learning classification algorithm originally designed for authorship verification (van Halteren, 2007), can also be used to classify speakers according to their gender, age, regional background and level of education by investigating the lexical content and the pronunciation of their speech. Our procedure differed from conventional procedures for speaker classification in that our algorithm analysed written representations of speech rather than the speech signal proper; it analysed orthographic and broad phonetic transcriptions of speech to identify regularities in the use of words and the pronunciation of speakers of different genders, ages, regional backgrounds and levels of education. These regularities were subsequently combined into feature sets: one set of features describing the use of words as reflected in the orthographic transcriptions, and a second set of features describing the pronunciation characteristics as reflected in the broad phonetic transcriptions. These feature sets were used to accept or reject unknown speakers as members of speaker classes that were defined in terms of the four aforementioned speaker characteristics. Since we wanted to study the performance of the algorithm with the individual features sets, the algorithm worked with one feature set at a time. The performance of the algorithm was evaluated through a comparison of its classification of unknown speakers with the information on the speakers as provided in the metadata of the speech material.

This chapter is organised as follows. In Section 5.2, we describe the corpus material and the transcriptions. In Section 5.3, we describe the classification algorithm, the definition of speaker classes, the two sets of classification features and our general experimental setup. Subsequently, in Section 5.4, we present and discuss the results of the classification experiments, and in Section 5.5 we present our conclusions and our plans for future research.

5.2 Corpus material and transcriptions

We conducted our classification experiments with transcriptions of spontaneous telephone dialogues from the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN), a 9-million word corpus comprising standard Dutch speech from native speakers in the Netherlands and in Flanders (Oostdijk, 2002). We considered recordings of telephone dialogues between speakers from the Netherlands only. These recordings were separated into two samples each (one sample per speaker). After excluding dialogues for which the metadata were incomplete as far as relevant for our classification variables (see Section 5.3.2), and after excluding samples of which large parts were tagged as unintelligible in the orthographic transcriptions, we counted 663 samples from 340 different speakers. These samples ranged from 321 to 2221 words in length and comprised a total of 689,021 word tokens.

In addition to the words in the orthographic transcriptions, we also considered their part-of-speech tags. The orthographic transcriptions of the words in the CGN were created fully manually, the part-of-speech tags were generated automatically and manually corrected afterwards (Oostdijk, 2002).

In order to study pronunciation characteristics we needed canonical representations of the words in the orthographic transcriptions (i.e. written representations of the standard pronunciation of the words in isolation from the context of neighbouring words (Laver, 1995)) and broad phonetic transcriptions reflecting their actual pronunciation in the speech recordings. We generated a canonical representation of each recording by substituting every word in the orthographic transcription with its representation in a canonical pronunciation lexicon. The broad phonetic transcriptions were generated automatically because the CGN provides manually verified phonetic transcriptions of only 115,574 out of the 689,021 words in our samples. We used an automatic transcription procedure which proved capable of closely approximating the manually verified phonetic transcriptions of the CGN (Van Bael et al., 2006; see also Chapter 3 of this thesis). In this procedure, the canonical representation of every utterance was first expanded into a network of alternative pronunciations. A continuous speech recogniser then chose the best matching phonetic transcription through forced recognition. In order to ensure the automatic generation of plausible phonetic transcriptions, we excluded speech utterances that, according to the orthographic transcriptions, contained non-speech, unintelligible speech, broken words and foreign speech. Samples containing overlapping speech were excluded as well. This resulted in automatic transcriptions for 252,274 out of 689,021 words, i.e. 136,700 words more than the 115,574 words for which the CGN could have provided a manually verified transcription.

5.3 Classification methodology

5.3.1 Classification algorithm

Linguistic Profiling is a supervised learning algorithm (van Halteren, 2007). It first registers all classification features (e.g. pronunciation processes) that occur in at least N training samples (e.g. speech samples of individual speakers) of a corpus[1]. The algorithm then builds a 'profile' of each training sample by listing the number of standard deviations the count of each of the classification features deviates from the average count in the whole corpus. Subsequently, class-specific profiles are generated by averaging the profiles of all training samples from a specific (speaker) class (e.g. male speakers). The distance between the profile of a test sample and the profile of a given class of speakers is compared with a threshold value in order to determine whether the speaker of the sample should be attributed to that speaker class. The degree to which the distance does or does not exceed the threshold value indicates the confidence of the decision. We evaluated the algorithm's classification accuracy by comparing its decisions with the actual characteristics of the test speakers as provided in the metadata of the CGN. Since Linguistic Profiling is a verification algorithm, we measured its accuracy initially in terms of False Accept Rates (FARs) and False Reject Rates (FRRs). Since these values are threshold-dependent, however, we present a threshold-independent derivative instead, viz. the Equal Error Rate (EER), which is the value at which the FAR and the FRR are equal.

5.3.2 Classification variables

We assigned the 663 selected samples to different classes according to the gender, age, regional background and level of education of the speakers.

The establishment of a male and a female speaker class was straightforward. We separated the samples into two classes: one class with 276 samples from 148 male speakers and another class with 387 samples from 192 female speakers.

All speakers were born between 1928 and 1981. We classified the speech samples age-wise according to two classification schemes. First, for every year, we generated a binary split of all speakers into those who were born in or before that year (e.g. $\leq$ 1955), and those who were born after that year (e.g. > 1955). This yielded classes with 24 to 639 samples from 11 to 329 speakers. In addition, for every year, we defined a class with subjects born within a symmetric eleven-year window around the target year (e.g. 1950 -1955- 1960). This yielded classes with 67 to 174 samples from 32 to 98 speakers.

[1] For each new classification task, the threshold (N) is empirically determined in order to keep the amount of information Linguistic Profiling has to deal with computable.

We retrieved the regional background of the speakers from the metadata of the CGN. We classified the speech samples in 16 classes according to the region speakers mainly lived in between the age of 4 and 16. Table 5.1 presents the distribution of samples per region. As a result of the large number of classes (we adhered to the original classification of the CGN), some classes contained only a few samples, in particular classes 2e (6), 3c (12) and 2f (13). However, since merging regional classes would probably have resulted in classes with more heterogeneous speech behaviour (which would probably have made speech from these classes harder to characterise and distinguish), we held on to the subdivision in 16 regional classes.

Table 5.1: Distribution of samples and speakers in terms of the speakers' regional backgrounds. From left to right: abbreviation, general geographical region in the Netherlands, specific geographical region, number of samples per class, number of individual speakers per class.

	general regions	specific geographical regions	sample	speaker
1a	central	South Holland, excl. Goeree Overflakee	105	55
1b		North Holland, excl. West Friesland	112	50
1c		West Utrecht, incl. the city of Utrecht	21	12
2a	transitional	Zeeland, incl. Goeree Overflakee + Zeeland Flanders	42	21
2b		East Utrecht, excl. the city of Utrecht	42	19
2c		Gelderland river area, incl. Arnhem + Nijmegen	52	27
2d		Veluwe up to the river IJssel	19	14
2e		West Friesland	6	4
2f		Polders	13	4
3a	peripheral, North East	Achterhoek	18	10
3b		Overijssel	37	20
3c		Drenthe	12	7
3d		Groningen	17	11
3e		Friesland	20	10
4a	peripheral, South	North Brabant	113	60
4b		Limburg	34	16
			663	340

The metadata of the CGN also provided us with information on the level of education of the speakers. The speakers were tagged as having enjoyed higher education (university or polytechnic), secondary education or only primary education (no completed secondary education). In our samples, we counted 256 speakers who had enjoyed higher education, 75 speakers with secondary education and only 9 speakers with primary education. Because of the skewness of the distribution of speakers in these three classes, and because we didn't have reason to believe that the 9 subjects of the third class would heavily increase the heterogeneity in the large second class if we would merge these classes, we merged the 9 speakers of the third class with the 75 speakers of the second class. As a result, two speaker classes were established: highly educated subjects (256 speakers in 496 samples) and moderately educated subjects (84 speakers in 167 samples).

5.3.3 Classification features

Per speaker class, the classification algorithm retrieved a set of lexical features from the orthographic transcriptions, and a set of pronunciation features from the broad phonetic transcriptions of the samples. The values of both feature sets were grouped into separate classification profiles modelling class-specific lexical use on the one hand and class-specific pronunciation characteristics on the other hand.

Lexical features

The lexical features largely resembled the features that were used for the authorship verification experiments in van Halteren (2007). This time, however, full syntactic analyses were not considered because the Amazon parser used in van Halteren (2007) has been developed for the analysis of written instead of spoken Dutch[2]. The lexical profiles represented the average utterance length in terms of number of word tokens, counts of uni-, bi-, and trigrams of words and the part-of-speech tags of the words. All counts were normalised for sample length by translating them to their frequency per 1000 tokens. In addition to these features, we tagged each utterance with information about the length, the linguistic status (declarative, interrogative and exclamatory, based on the punctuation marks) and the speaker (current speaker or interlocutor) of the preceding utterance. Only the features occurring in at least five samples were used. This led to a feature set of about 150.000 features potentially useful for classification.

Pronunciation features

We characterised 'pronunciation features' in terms of the segmental differences between the canonical (standard) representations and the broad phonetic transcriptions of the words in the speech samples. We aligned the canonical and broad phonetic transcriptions with ADAPT, a dynamic programming algorithm designed to align strings of phonetic symbols according to their articulatory distance (Elffers et al., 2005). Figure 5.1 illustrates the alignment of a canonical (Can) and a broad phonetic transcription (PT), and the derivation of a pronunciation process: the deletion of schwa.

The segmental differences between a canonical and a broad phonetic transcription can be influenced by (at least) four main variables: the socio-situational setting (transcriptions of spontaneous speech typically yield more mismatches than transcriptions of prepared speech (Binnenpoorte, 2006), the consistency of the transcriptions (human transcribers may not always transcribe in a consistent manner (Cucchiarini, 1993), the use of words (the use of

[2] Part of the CGN is annotated for syntactic structure, but the amount of annotated data would have been insufficient to be of use for our experiments.

specific words triggers specific pronunciation processes), and the pronunciation habits of the speakers (the focus of our study). Since we aimed at classifying speakers according to genuine speaker-specific pronunciation features only, we tried to filter out all pronunciation features that were due to the other three variables.

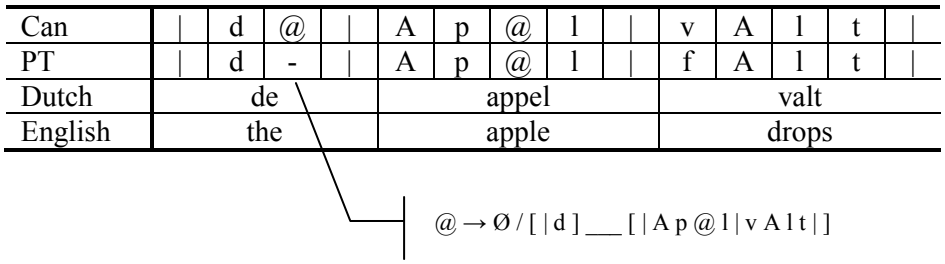

Figure 5.1: Alignment of a canonical (Can) and a broad phonetic transcription (PT) and derivation of a pronunciation process (deletion of schwa). SAMPA symbols are used, word boundaries are marked as vertical bars.

The first two variables (the socio-situational setting and the consistency of the transcriptions) were irrelevant for our study since we considered the transcriptions of speech uttered in one socio-situational setting only, and since the transcriptions were generated by a consistent automatic transcription procedure. This left us with one more variable to control: the lexical context in which the pronunciation processes occurred, which we modelled by means of the frequency of the current word and information about its context (its co-occurrence with surrounding words and the position of the word in the utterance).

We controlled for lexical context by means of a three-step procedure. First, we set up a ten-fold cross-validation training in which we consecutively built ten models for the impact of the lexical context on pronunciation, each time on the basis of 90% of the samples. The models represented the counts of all pronunciation processes observed in their canonical contexts. Next, we used each of these models in turn to predict the pronunciation in the left-out samples. For every canonical phone the pronunciation model predicted the probability of different phones being actually pronounced, considering all canonical contexts seen in the training material. As a final step, we compared the predicted pronunciation processes with the pronunciation processes observed in the automatic phonetic transcription. To this end, we counted the actual occurrences of all pronunciation processes in every sample, and for each process we calculated the difference between the predicted and the observed probability. These differences were considered to mainly indicate speaker-specific pronunciation processes, since these pronunciation processes were present *in addition to* the pronunciation processes that were predicted on the basis of the lexical context of the pronunciation processes. This additional variation was numerically represented as a feature vector of 94 numerical values, one for each of the 94 different pronunciation processes that were encountered in our material.

To investigate to what extent our approach was successful in removing the influence of the speakers' use of words on the observed pronunciation processes, we computed the Kullback-Leibler distance between the predicted and the observed pronunciation processes. This distance halved when the predictions were based on the observations of pronunciation processes in their lexical context instead of on the observations of the pronunciation processes by themselves, without considering their context. This leads us to believe that a significant part of the influence of the lexical context was indeed modelled by our method.

5.3.4 Experimental setup

We organised our classification experiments as ten-fold cross-validations. To this end, we divided the 663 samples in ten mutually exclusive sample sets of comparable size. Each speaker occurred in one set only. Per classification variable (gender, age, regional background and level of education) and per feature set (lexical and pronunciation), we consecutively used nine sample sets to train the algorithm, and the remaining set to test the algorithm. Each time, Linguistic Profiling was trained and tested with a range of parameter settings. Upon completion, we considered the algorithm's accuracy at the parameter settings yielding the best performance over all ten folds in order to determine its performance ceiling.

After running our experiments, it became clear that the use of this Oracle approach had a negative consequence, in particular when we assessed the algorithm's performance for speaker classes with a small number of samples. We found that the EERs at the best performing parameter settings were lower than 50%, even when we attempted the classification of speakers in classes with randomly selected speakers. This is not surprising: there will always be some degree of variance around the expected value of 50% accuracy, and by selecting the best performing settings we are likely to end up with a score better than 50%. This effect grows stronger as the number of samples in the classification profiles becomes smaller.

In order to determine whether the algorithm's classification was above or below chance rate, we experimentally determined the mean and standard deviation of the algorithm's EER for the classification of speakers in *randomly* selected speaker classes of various sizes. When 300 or more random samples were used, we found a mean *random group EER* of 44% with a standard deviation under 2%. When our algorithm considered 50 to 100 random samples, we found a mean random group EER of 40% with a standard deviation of 3%. When smaller groups of random samples were considered, the mean random group EER gradually decreased while its standard deviation increased.

To facilitate the interpretation of the classification results in the upcoming sections, we compare each EER with the expected distribution of the random group EERs. We mark each EER with one asterisk if the probability that it belongs to the distribution of the random group EERs is smaller than 0.05, with two asterisks when $p<0.01$ and with three asterisks when $p<0.001$. In all cases where $p<0.05$, we will call the classification "*effective*". Since both the expected EER value and the variance depend on group size, all values reported below for different speaker classes can only be compared directly if the number of speech samples in the classes is comparable.

5.4 Classification results

5.4.1 Classification in terms of gender

For both genders, we conducted a ten-fold cross-validation in which nine tenths of the transcriptions were used to identify gender-specific lexical and pronunciation features, and in which the transcriptions of the remaining samples were used to test the classification algorithm. Table 5.2 presents the results of this experiment for both genders and feature sets. The results were obtained with the algorithm's optimal parameter settings for each of the two feature sets.

Table 5.2: Best possible speaker classification in terms of gender with lexical and pronunciation features.

gender	# samples	lexical features	pronunciation features
		EER (%)	EER (%)
male	276	23 ***	41
female	387	24 ***	42

Whereas the use of the lexical features led to an effective classification with error rates of about 24%, the use of the pronunciation features did not. In other words, the pronunciation features could not help the algorithm distinguish between the phonetic transcriptions of male and female speakers. The frequent misclassification of speakers from their pronunciation features may be due to several reasons. The most obvious reason would be the absence of gender-specific pronunciation characteristics at the broad phonetic level. A more disturbing reason (disturbing because it would question the validity of our automatic phonetic transcriptions as a knowledge source for our experiments), would be an inadequate representation of gender-specific pronunciation features in the automatic phonetic transcriptions.

There are two reasons to assume that the mediocre classification performance of our algorithm was due to the absence of gender-specific pronunciation characteristics at the broad phonetic level rather than to inadequacies in the automatic phonetic transcriptions. First, the linguistic literature has not yet reported systematic gender-specific pronunciation differences at the broad phonetic level. The only systematic gender-specific pronunciation characteristics that have so far been reported were based on measurements of the overall speech rate (Byrd, 1994; Verhoeven et al., 2004), and on measurements at levels of finer phonetic detail (e.g. a structural difference between the dimensions of the vowel space of male and female speakers (Henton, 1994)). None of these gender-specific pronunciation characteristics can be reflected in a broad phonetic transcription of speech, e.g. in the form of systematic phone deletions or substitutions. Second, our results are in line with Binnenpoorte et al. (2005), who could not discover gender-specific pronunciation characteristics through the alignment of a canonical and a manually verified (instead of an automatic) transcription of male and female speech from the CGN either.

5.4.2 Classification in terms of age

For every year of birth between 1928 and 1981, we tried to classify the speakers in terms of them being born before that year (24 to 631 samples per class), after that year (32 to 639 samples per class), or in an eleven-year window around that year (67 to 174 samples per class - see Section 5.3.2). Since these classification experiments yielded many data points, we confine ourselves to a description of the general tendencies.

Despite the fact that the algorithm was able to retrieve and successfully use age-specific pronunciation features for most of these classes, it still performed better with the lexical profiles than with the pronunciation profiles. The binary before/after classifications showed relatively stable classification accuracies: ignoring three outliers at each side of the time scale, the EERs ranged between 18% and 23% for the lexical profiles (with a mean EER over all age classes of 20.5%) and between 26% and 36% for the pronunciation profiles (mean EER over all classes: 32.4%). The use of the lexical profiles consistently led to effective classification ($p<0.001$), the use of the pronunciation profiles as well ($p<0.01$, and in 90% of the tests even $p<0.001$).

The classification of speakers according to the eleven-year intervals showed more variation: ignoring the same three outliers at each side of the time scale, we obtained error rates between 19% and 41% with the lexical profiles (mean EER over all classes: 32.0%), and between 28% and 46% with the pronunciation profiles (mean EER over all classes: 38.5%). The use of the lexical profiles led to effective classification for the years at the outskirts of the time scale ($p<0.001$ for the years between 1928 and 1942, and between 1973 and 1981) whereas there was hardly any effective classification noticeable for the years between 1942 and 1973. The use of the pronunciation features showed a similar pattern, although fewer effective classifications were found.

5.4.3 Classification in terms of regional background

Table 5.3 presents the results of the classification of our speakers according to the regional background they lived in between the age of 4 and 16. We classified the speakers in terms of 16 geographical regions (see Table 5.1 in Section 5.3.2) and by means of the two feature sets.

Table 5.3 shows that the classification algorithm obtained effective classification for 10 out of 16 regions when using the lexical classification features. This indicates that the orthographic transcriptions of (at least part of) the investigated speech contained useful information with which our classification algorithm could classify unknown speakers. The EERs in Table 5.3 can only be compared for speaker classes comprising a comparable number of speech samples, because the EERs decreased when speaker classes with fewer samples were considered (e.g. compare the 36% EER with class 1b, which was made up of 112 samples, with the 26% EER with class 1c with only 21 samples). This means that we cannot draw conclusions about specific regions being more easily recognised than other regions.

Table 5.3: Best possible speaker classification in terms of regional background with lexical and pronunciation features.

region	# samples	lexical features	pronunciation features
		EER (%)	EER (%)
1a	105	35 *	38
1b	112	36 *	40
1c	21	26 *	29
2a	42	23 ***	34
2b	42	36	37
2c	52	32 *	40
2d	19	30	32
2e	6	21	21
2f	13	14 **	30
3a	18	19 **	33
3b	37	27 *	37
3c	12	23	24
3d	17	36	26
3e	20	38	31
4a	113	32 ***	40
4b	34	27 *	22 ***

The results in Table 5.3 also show the inability of the algorithm to retrieve and use geographically determined pronunciation features from the broad phonetic transcriptions. The classification algorithm was only able to effectively classify speakers of one region (Limburg, a peripheral region in the South East of the Netherlands).

The poor performance of the classification algorithm with the pronunciation features can be due to several reasons. First of all, some of the above mentioned regions may have characteristic features, but of a kind that are usually not represented at the broad phonetic level. For example, the Dutch phoneme /r/ has many allophonic variants, some of which have been reported characteristic for specific regions in the Netherlands (Verstraeten and Van de Velde, 2001). However, in our study these different realisations could not be used for classification because the broad phonetic transcriptions did not distinguish allophonic variants of the phoneme /r/. A second possible explanation for the disappointing performance of our algorithm is the absence of distinguishing pronunciation features in our automatic phonetic transcriptions. To verify whether this could indeed be so, we examined the pronunciation of word-final /n/ preceded by schwa in plural nouns and verbs, since this pronunciation process is known to be typical for speakers in specific regions of the Netherlands, notably 2d and 3a (Hol, 2006). A comparison between the speech samples for which both automatic and manually verified phonetic transcriptions were available showed that the automatic phonetic

transcriptions did not represent the pronunciation of such word-final /n/s, whereas the manually verified phonetic transcriptions of the CGN did at least in some cases. A third possible explanation for the mostly ineffective classification performance is the potential mismatch between the geographical boundaries of the 16 regions defined in the CGN and the regions that can actually be characterised by means of outspoken pronunciation features. A fourth potential explanation is the heterogeneity of the speaker populations in the regional classes, either because the pronunciation features in these classes are inherently heterogeneous or because some speakers in the CGN are not particularly representative of their region. Finally, of course, we should also consider potential limitations of Linguistic Profiling for the purpose of classifying speakers on the basis of pronunciation features. Perhaps its capabilities were hampered by the fact that it could only use 94 pronunciation features, while there were some 150.000 lexical features, comparable to the number of features used in van Halteren (2006).

Further research is needed to clarify the way in which the above mentioned factors affect the classification of speakers on the basis of manual or automatic broad phonetic transcriptions.

5.4.4 Classification in terms of education level

Finally, we investigated whether our classification algorithm was able to classify speakers in terms of their level of education. Table 5.4 presents the classification results for the two speaker classes (highly educated and moderately educated) with both the lexical and the pronunciation features. Again, we show the Equal Error Rates at the algorithm's optimal parameter settings.

Table 5.4: Best possible speaker classification in terms of education level with lexical and pronunciation features.

level of education	# samples	lexical features	pronunciation features
		EER (%)	EER (%)
highly educated	496	41	46
moderately educated	167	41	44

The results in Table 5.4 show that the algorithm was not able to classify speakers effectively in terms of their level of education. The classification results with the pronunciation features reflect the inconclusive results reported in Keune et al. (2005). While they found significant differences between the reductions of phones in 14 frequent words ending in –lijk spoken by highly versus moderately educated Flemish speakers (the moderately educated speakers reduced more phones), there was no significant difference between the phone reductions of highly and moderately educated speakers from the Netherlands.

Although our results do not imply that speakers cannot be categorised according to the influence of their education on their speech, the high EERs do imply that the lexical features

as well as the pronunciation features were unsuitable for classifying speakers according to their education level. Future research should clarify whether a further division of the speakers into smaller, more specific classes can improve classification accuracy.

5.4.5 More specific speaker classes

In the previous sections, we classified speakers in classes that were defined by one speaker characteristic (gender, age…) at a time. However, someone's speech is likely to be influenced by the interplay of all four aforementioned speaker characteristics. This implies that, when we classify speakers in broad classes of which all members have only one characteristic in common, the 'class-specific' speech features may show a great deal of dispersion. Evidently, speaker classification with very broad and therefore perhaps partially overlapping classification profiles for different speaker classes is more difficult than speaker classification with well defined and more exclusive classification profiles.

Therefore, we attempted an integrated classification of our speakers according to all four speaker characteristics by using classes of speakers for which all four characteristics were fixed. In order to have sufficient training data for each combined class, we restricted this experiment to the classification of highly educated women who were born before or in 1975 and who were raised in region 1a (South Holland) or 4a (North Brabant), and the classification of highly educated women who were born after 1975 and who were raised in North Brabant. Class profiles were created for each of these classes. Table 5.5 presents the results of this classification experiment.

Table 5.5: Best possible speaker classification in terms of three specific speaker classes according to a joint assessment of four speaker characteristics: gender, age, education, regional background.

highly educated women		# samples	lexical features	pronunciation features
born	raised in		EER (%)	EER (%)
≤1975	1a	29	24 **	36
≤ 1975	4a	28	30	30
> 1975	4a	23	31	28

In order to evaluate the possible benefit of classifying speakers in more specific speaker classes rather than in general classes, one would ideally want to compare the EERs in Table 5.5 with the EERs reported in Sections 5.4.1 to 5.4.4. However, as was explained in Section 5.3.4, such a direct comparison is impossible because of the different number of samples in the speaker classes in the previous sections. It is possible, however, to compare the EERs obtained with the lexical and the pronunciation features for the three specific speaker classes in Table 5.5. These comparisons (24-36%, 30-30%, 31-28%) show that per speaker class, the EERs obtained with both feature types were much more similar than in the previous sections.

We hypothesise that in the previous sections, where we classified speakers in general speaker classes that were defined by only one common speaker characteristic (e.g. gender), classification was affected by an influence from the interplay of the remaining speaker characteristics (age, regional background and level of education) on the speech features in the classes. It may well be that in these circumstances, the classification algorithm could still benefit from the abundance of lexical classification features (around 150.000) to use the most distinguishing features for classification and to ignore less characteristic features. At the same time, the algorithm may have had more difficulties to select and use features out of the much smaller set of 94 pronunciation features which were characteristic of the classes and which were not influenced by an interplay of speaker characteristics.

5.5 Conclusions and plans for future research

We investigated whether Linguistic Profiling, a supervised learning algorithm originally designed for authorship verification, can be used to classify speakers according to their gender, age, regional background and level of education on the basis of the lexical content and the pronunciation of their speech. Our approach differed from conventional speaker classification procedures in that our algorithm analysed written representations of speech rather than the speech signal proper; it analysed orthographic and broad phonetic transcriptions of speech in order to identify regularities in lexical content and pronunciation.

We conducted experiments to determine the performance of our algorithm for speaker classification with the aforementioned lexical and pronunciation features. These experiments showed that the algorithm was often able to retrieve and use characteristic lexical features from the orthographic transcriptions. The lexical features enabled the classification algorithm to distinguish between male and female speakers, to classify speakers in terms of their age, and to determine the region speakers spent most of their childhood in (this held for 10 out of 16 investigated regions). Despite these encouraging results, however, the use of the lexical features proved insufficient to effectively classify speech from moderately or highly educated speakers and from people who spent their childhood in specific (6 out of 16 investigated) regions in the Netherlands. Moreover, the algorithm's performance is probably not good enough for operational speaker classification: in general, we found equal error rates between 20% and 40%.

When the classification algorithm had access only to the pronunciation features as reflected in our automatic broad phonetic transcriptions, it was hardly ever able to classify speakers effectively. We have argued that this may be explained by 1) the absence in the material of distinguishing pronunciation features at the broad phonetic level, 2) the failure of the automatic phonetic transcription procedure to capture distinguishing pronunciation features, 3) a mismatch between our speaker classes and groups of speakers that possibly show distinguishing speech features, 4) the heterogeneity of our speaker classes (either because they are inherently heterogeneous or because the speakers were not representative of their classes), and 5) the limitations of our algorithm for classification with a small number of classification features.

In future research, some of these potential explanations may be further investigated. As for 1), we had hoped that the relatively large amounts of broad phonetic transcriptions would enable our algorithm to identify class-specific pronunciation features at the broad phonetic level. However, our approach to defining potentially useful pronunciation processes resulted in fewer than 100 such features, which appeared insufficient to distinguish speaker classes effectively. It remains to be seen if and how the number and the distinctiveness of the pronunciation features can be increased. One option might be to move towards more detailed phonetic transcriptions. This would increase the number of possible mappings between canonical representations and actual realisations, and hence potentially also the number of different pronunciation processes that can be used for classification. This approach may seem counterproductive because it might reduce the number of pronunciation processes that occur at least five times (the criterion used in this study). However, if more detailed transcriptions can be made reliably, we might gain after all, since the use of more diverse phonetic symbols can result in the definition of more diverse but also more systematic phone mappings representing characteristic pronunciation features. At the same time it is clear that the further we would move away from a broad phonetic transcription of speech, the closer we would come to traditional signal-based classification procedures. As for 2), we have identified at least one regional pronunciation phenomenon, viz. the presence of word-final /n/ preceded by schwa in plural nouns and verbs, which was not systematically represented in the automatic transcriptions. It may well be that the same holds for other pronunciation phenomena that are conventionally considered as characteristic for some geographical region; the automatic transcription procedure, which was based exclusively on local properties of the speech signal may have selected its symbols less 'systematically' than the human transcribers who may have been biased towards conventional regional characteristics on the basis of subtle cues in the signal. Again, this seems to suggest that we should try and move towards more detailed phonetic transcriptions. As for 3) and 4), we may attempt to classify speakers in more specific classes, hopefully with more homogeneous speech behaviour. In most cases, this is likely to mean a subdivision of the classes used in this study. Recall that we classified our speakers in just 16 predefined geographical regions, and that we attempted the classification of speakers in just two classes defined by their level of education. The training and use of more specific speaker classes may increase the homogeneity of speech characteristics in these classes, but it would inevitably also introduce a data sparseness problem. Finally, as for 5), we may consider increasing the number of classification features for our algorithm, but we have already argued that it is not obvious how this can be accomplished. Alternatively, we may consider investigating classification techniques that are designed to operate with smaller numbers of features.

Finally, for a real application rather than for a scientific investigation like this study, it will probably be suboptimal to base classifications on a single type of classification features. For the best possible classification, we should give the classifier access to as many and as large a variety of features as possible. This means combining both the lexical and pronunciation features presented here, and probably also other features which have proven useful for speaker classification, e.g. acoustic features that can be directly retrieved from the speech signal as illustrated in Müller and Schötz (2007).

CHAPTER 6
GENERAL DISCUSSION

The studies in this dissertation were conducted with three aims in mind. The first aim was to reconsider the general applicability of validation procedures that validate phonetic transcriptions through a comparison with a human-made reference transcription (Chapter 2). The second aim was to investigate to what extent one can approximate human-like transcriptions with a fully automatic transcription procedure (Chapter 3). The third aim was to investigate the usefulness of both semi-automatic (manually verified) and fully automatic phonetic transcriptions as tools for the disclosure of linguistic knowledge in large speech corpora (Chapters 4 and 5). In this chapter, I discuss the results of the studies reported in Chapters 2 to 5.

6.1 Validation of broad phonetic transcriptions

Large speech corpora such as the SpeechDat databases are usually validated through a systematic test to verify that they meet their specifications. Corpus specifications (e.g. the intended number of female speakers) and tolerance margins for these specifications (e.g. a permitted deviation of 5%) are normally defined before or in the initial stage of corpus compilation. Speech corpora are validated to assure their quality and/or to improve their quality on the basis of the suggestions made by the validating authority. Validation reports are sometimes included with corpora to provide an indication of the extent to which they meet their specifications (Van den Heuvel and Sanders, 2006).

The validation of speech corpora also involves the validation of their transcriptions. The validation of phonetic transcriptions, whether provided in a pronunciation lexicon or as a separate annotation layer, is usually conducted by an external expert (a *validator*) who assesses the acceptability of a representative subset of the transcriptions in terms of their correspondence to the transcription guidelines (e.g. Goddijn and Binnenpoorte, 2003; Pitt et al., 2005). In order to avoid pointless discussions due to subjective judgements, validators are instructed to give the provided transcription the benefit of the doubt and to consider a transcription erroneous only if it is unmistakably linguistically implausible (if transcriptions in lexica are assessed) or if it clearly does not match the actual speech signal (if time-aligned transcriptions are assessed). In the latter case, one of the usual formal validation procedures for phonetic transcriptions consists of a comparison of the sequence of symbols in the corpus transcription with the symbols in a reference transcription. If the number of deviations is within the tolerance margin defined in the corpus specification, phonetic transcriptions are considered valid. The fewer deviations between a transcription and a reference transcription, the better the transcription is considered to be.

In Chapter 2, we reconsidered the general applicability of this generic validation procedure. The validation of phonetic transcriptions in terms of their deviations from a reference transcription is ignorant of the purpose(s) transcriptions are generated for. Since phonetic transcriptions are often used for training automatic speech recognition systems, and since the relation between recognition performance and the quality of such transcriptions expressed as the deviation from a reference transcription is difficult to predict, we tested to

what extent this relation holds. The purpose of this study was to investigate whether the conventional validation method offers a useful indication of a transcription's suitability for the training of an automatic speech recognition system. We compared a basic canonical transcription and a manually verified phonetic transcription in terms of their resemblance to a consensus transcription (Shriberg et al., 1984) and in terms of the quality measure of the application we would use them for, viz. automatic speech recognition. The relation between the results of the conventional and the application-oriented validation procedure was not straightforward.

The outcome of the traditional validation method was clear: the quality of the manually verified phonetic transcriptions was assessed as much higher than the quality of the canonical representations, because the manually verified phonetic transcriptions differed less from the reference transcription than the canonical representations. The application-oriented validation method gave another estimate of the transcriptions: the use of manually verified phonetic transcriptions and canonical representations did not yield significantly different recognition performance. This implies that the manually verified phonetic transcriptions and the canonical representations were equally suitable for the purpose of developing an automatic speech recognition system.

A comparison of the outcomes of the two validation methods supports different conclusions. First of all, it should be stressed that the application-oriented validation method did not contradict the usefulness of manually verified phonetic transcriptions for the development of speech recognition systems, since we did not get better recognition results with the canonical representations. As a matter of fact, for other purposes than training today's automatic speech recognition systems, the story may well be different. For many, if not all, research in phonetics, to name just one example, it will remain essential for transcriptions to reflect the speech signal as closely as possible. For such purposes, manually verified phonetic transcriptions should definitely be preferred over canonical representations, for canonical representations do not (or only partially) represent the pronunciation variation observed in everyday speech.

A more important conclusion, however, is that the traditional validation method unnecessarily favoured the manually verified phonetic transcriptions above the canonical representations which can be obtained with substantially less effort. This was not justified by the outcome of our recognition experiment; the use of the canonical representations yielded similar recognition results. Most probably, this finding can be explained by a specific feature of statistical pattern recognition. The use of canonical representations both for training and testing results in a smaller training-test mismatch than using manual transcriptions in training and canonical representations in testing.

To sum up, we found no consistent relationship between the distance of a broad phonetic transcription to a reference transcription on the one hand, and the influence of that transcription on the recognition performance of a continuous speech recogniser on the other hand. This outcome has two implications. First of all, our experiments showed that canonical transcriptions can serve the purpose of training automatic speech recognition systems equally well as more expensive manually verified phonetic transcriptions. This would never be

reflected in a traditional validation of transcriptions. Therefore, developers of automatic speech recognition systems need to invest in the manual verification of phonetic transcriptions only for specific purposes, e.g. in-depth analysis of what caused recognition errors, or as a tool for modelling pronunciation variation. Second, and most importantly, the results of our experiments imply that phonetic transcriptions should preferably be validated in terms of the application they will serve because a higher resemblance to a purpose-independent reference transcription proved no guarantee for a transcription to be better suited for the development of automatic speech recognition systems.

This is a shift in paradigm with important implications for the future. Our results suggest that a manual reference transcription should no longer be considered the self-evident golden criterion to generate phonetic transcriptions with regardless of the purposes for which the transcriptions will be used. Rather, our results suggest that the validity of transcriptions should be tested by using the criterion measure for the application. By taking this approach the status of human reference transcriptions shifts from them being a golden standard and final goal for any transcription to just another competing transcription that can be outperformed by any other transcription that scores better for a given task. This is a new perspective for transcription validation that deserves exploration for other applications than automatic speech recognition as well.

6.2 Automatic generation of broad phonetic transcriptions

Large speech corpora such as the SpeechDat databases are invariably delivered with a pronunciation lexicon that contains a canonical transcription of the words in the orthographic transcription (Van den Heuvel et al., 2001). Such a lexicon can be used for generating a hypothetical ‘canonical’ phonetic transcription from the orthography by substituting every word with its canonical transcription. In addition, money and time permitting, some speech corpora are at least partially provided with a manually verified broad phonetic transcription. Such transcriptions are made through a joint effort of man and machine: human experts check and correct a (usually canonical) example transcription according to a transcription protocol and their own perception of the speech signal. Parts of Switchboard (Greenberg, 1997) and the Spoken Dutch Corpus (Goddijn and Binnenpoorte, 2003) were transcribed this way.

Over the years, manually verified phonetic transcriptions have proven useful for diverse purposes such as lexical pronunciation variation modelling for automatic speech recognition (ASR - Strik, 2001), unit selection for speech synthesis (Mizutani and Kagoshima, 2005), automatic pronunciation training and assessment in Computer Assisted Language Learning and general research on pronunciation variation (Greenberg et al, 1996; Riley et al., 1999). However, since the employment of human transcribers for such verification procedures is time-consuming and expensive, we investigated to which extent one can approximate manually verified transcriptions by means of quicker and cheaper automatic transcription procedures. In Chapter 3, we tried to approximate manually verified phonetic transcriptions of

read speech and spontaneous telephone dialogues from the Spoken Dutch Corpus (Oostdijk, 2002) by means of a diverse set of automatic transcription procedures.

One transcription procedure proved capable of approximating the phonetic transcriptions of the Spoken Dutch Corpus quite well. This procedure requires the use of a small seed corpus with manually verified target transcriptions, an orthographic transcription of the corpus to be transcribed, a canonical pronunciation lexicon, a basic speech recognition system and software for the implementation of decision trees. The procedure hinges on the availability of powerful machine learning techniques, such as decision trees. We first used decision trees for learning the (statistics of the) discrepancies between the phonetic symbols in a canonical 'base' transcription and the manual target transcriptions. The canonical transcription was generated by replacing every word in the orthography with its canonical pronunciation taken from the lexicon. Subsequently, we used the trained decision trees for expanding a similar base transcription of the remainder of the corpus into a pronunciation network with multiple pronunciation variants per word. The automatic speech recognition system chose the best matching pronunciation variant through forced recognition.

Our target transcriptions, viz. the manually verified phonetic transcriptions of the Spoken Dutch Corpus, were based on a canonical example transcription. This probably explains why our transcription procedure approximated the target transcriptions best when starting from a canonical transcription instead of a transcription which was based on the speech signal or on phonological knowledge reported in the literature. This way the forced recognition (in combination with the decision trees for generating pronunciation variants) performed the same task as the human transcribers of the Spoken Dutch Corpus: they modified the canonical example transcriptions in the direction of the human perception of the speech signal.

One should bear in mind that a canonical point of departure may be suboptimal to approximate target transcriptions that are based on a different type of example transcription. However, the machine learning procedure on which our approach is based seems sufficiently general and powerful to approximate different types of transcriptions, as long as learning can be initialised from a base transcription that is not too far from the target transcription. The larger the distance between the base transcription and the target transcription, the more discrepancies have to be modelled. The more complex the discrepancies, the more pronunciation variants (with smaller lexical probabilities) will be generated for forced recognition. Obviously, the larger the number of pronunciation variants and the smaller the difference between their lexical probabilities, the more difficult it is for a speech recognition system to select the best matching pronunciation variant.

The findings of this study may be beneficial for future corpus designers, because the use of machine learning techniques may offer a quick and cheap alternative to employing human transcribers for the manual verification of example transcriptions. Our study suggests that it is sufficient to verify the phonetic transcription of only a small portion of a corpus by hand in order to automatically generate similar transcriptions for the remainder of the corpus by means of automatic procedures. Even though our procedure requires the availability of a small seed corpus of target transcriptions, the generation of such a small set of transcriptions is almost always (much) cheaper than providing manual transcriptions for the full corpus.

6.3 Research with broad phonetic transcriptions

6.3.1 Linguistic research with manually verified phonetic transcriptions

In Chapter 4, we studied the frequencies of phone and syllable deletions in the manually verified phonetic transcriptions of the spontaneous telephone conversations from the Spoken Dutch Corpus. We found average phone and syllable deletion rates of 7.57% and 5.46% respectively. 22.76% of the content words and 18.1% of the function words had at least one phone missing, and 6.66% of the content words and 7.10% of the function words had at least one syllable missing. These statistics indicate that phones and syllables are frequently deleted in everyday conversational Dutch.

In addition to studying the statistics of phone and syllable deletions, we investigated to what extent various (socio)linguistic factors represented in the parallel annotations and the metadata of the Spoken Dutch Corpus are predictive for the deletion of phones and syllables. We did this by fitting two separate models to the data: one model for phone deletion and one for syllable deletion. Since we aimed at studying the simultaneous effects of various factors on deletion rather than the effects of a small number of selected factors at a time, we made use of mixed-effect models. Such statistical models are able to include both fixed-effects and random-effects factors, and they are convenient for studying complex corpus data because they can cope with missing data, because they can model the effects of many factors in one model, and because they are parsimonious with regard to the number of parameters they need to estimate for their random and fixed-effects factors. Until recently, however, two or more random-effects factors in the same mixed-effects model could only be modelled as nested. This imposed serious limitations on the use of mixed-effects models for linguistic studies.

The recent possibility to include random-effects factors as crossed instead of nested now makes it possible to include several random-effects simultaneously as independent sources of random variation. This no longer presupposes the assumption that the levels of random-effects factors are dependent of each other, and this makes it possible to assess in a methodologically sound way in one model which linguistic and sociolinguistic variables predict deletions over and above the random variation that comes with the sampled levels of the random effects. Including random effects as crossed instead of nested decreases the risk of type I errors because it prevents fixed-effects factors of modelling random variation that should be better modelled by means of by-subject and by-item adjustments (i.e. by means of random-effects factors). Through this new modelling technique, linguistic phenomena such as phone and syllable deletion can now be studied in a methodologically sound way in corpus data as a function of the interplay of many factors instead of in controlled experimental environments designed for studying the effects of one or a few factors at a time.

We found that our mixed effects models for phone and syllable deletion had several effects in common, which implies that both types of deletion are to a large extent influenced by the same factors. The strongest factors across both models were lexical stress, word duration and whether the following word starts with a vowel or a consonant.

Even though the collection of factors included in our study is not exhaustive (e.g. in future research we could include signal-based information and correlations between factors), our study illustrates new opportunities for analysing rich and complex corpus data by means of generalised linear mixed effects models with crossed random effects. The use of such statistical models is useful for exploratory research like ours (we investigated the simultaneous effects of a wide variety of linguistic and sociolinguistic factors on segment deletion) as well as for hypothesis testing (e.g. for testing whether a specific factor has an effect on a specific speech process in addition to the effects of other factors, or for testing whether a correlation between two factors has a larger effect than another correlation between factors). Since mixed effects models with crossed random effects can be fitted to all kinds of data with random and fixed effects, this technique offers new promising opportunities for many kinds of linguistic studies in large and richly annotated speech corpora. Considering the wide applicability of mixed effects models, these studies can easily transcend the field of pronunciation variation modelling.

The findings presented in Chapter 4 are clearly affected by the choice for the reference transcription relative to which we identified the deletions. As in previous research for (American) English the reference transcriptions consisted of concatenations of citation forms of the words in an accurate verbatim transcription of the speech. Such a reference does not account for (almost) obligatory assimilations and degeminations across word boundaries. Future research should clarify the impact of this decision on the general trends in deletions of phones, but also of syllables.

6.3.2 Linguistic research with automatic phonetic transcriptions

In chapter 5, we used automatic phonetic transcriptions for automatic speaker classification. We used a classification algorithm (Linguistic Profiling; van Halteren, 2004) which had previously proven successful for authorship verification, to automatically learn characteristic speech habits from broad phonetic and orthographic transcriptions of speech from speakers whose gender, age, level of education and regional background were known. This knowledge was subsequently used to determine the gender, age, level of education and regional background of unknown speakers on the basis of the orthographic and broad phonetic transcription of their speech. In order to train the classification algorithm, we required more phonetic transcriptions than were manually verified in the Spoken Dutch Corpus. Therefore, we used the most optimal transcription procedure from Chapter 3 to automatically generate a phonetic transcription of all speech that fitted our experimental design. Our approach clearly differed from conventional speaker classification procedures in that our algorithm analysed written representations of speech rather than the speech signal proper.

Our experiments showed that the classification algorithm was often able to retrieve and use characteristic lexical features from the orthographic transcriptions. The lexical features enabled the classification algorithm to distinguish between male and female speakers, to classify speakers in terms of their age, and to determine the region speakers spent most of

their childhood in (this held for 10 out of 16 investigated regions). Despite these encouraging results, however, the use of the lexical features proved insufficient for the algorithm to effectively distinguish between speech from moderately and highly educated speakers. Moreover, the algorithm's performance is probably not good enough for operational speaker classification: in general, we found equal error rates between 20% and 40%.

When the classification algorithm had access only to the pronunciation features as reflected in our automatic broad phonetic transcription, it was hardly ever able to classify speakers effectively. We have argued that this may be explained by 1) the absence in the material of distinguishing pronunciation features at the broad phonetic level, 2) the failure of the automatic phonetic transcription procedure to capture distinguishing pronunciation features, 3) a mismatch between our speaker classes and groups of speakers that possibly show distinguishing speech features, 4) the heterogeneity of our speaker classes (either because they are inherently heterogeneous or because the speakers were not representative of their classes), and 5) the limitations of our algorithm for classification with a small number of classification features.

In future research, some of these potential explanations may be further investigated. As for 1), we had hoped that the relatively large amounts of broad phonetic transcriptions would enable our algorithm to identify class-specific pronunciation features at the broad phonetic level. However, our approach to defining potentially useful pronunciation processes resulted in fewer than 100 such features, which appeared insufficient to distinguish speaker classes effectively. It remains to be seen if and how the number and the distinctiveness of the pronunciation features can be increased. One option might be to move towards more detailed phonetic transcriptions. This would increase the number of possible mappings between canonical representations and actual realisations, and hence potentially also the number of different pronunciation processes that can be used for classification. If more detailed transcriptions can be made reliably, we might gain after all, since the use of a larger number of different phonetic symbols can result in the definition of more diverse, but also more systematic phone mappings representing characteristic pronunciation features. At the same time it is clear that the further we would move away from a broad phonetic transcription of speech, the closer we would come to the signal-based classification procedures. As for 2), we have identified at least one regional pronunciation phenomenon, viz. the presence of word-final /n/ preceded by schwa in plural nouns and verbs, which was not systematically represented in the automatic transcriptions. It may well be that the same holds for other pronunciation phenomena that are conventionally considered as characteristic for some geographical region; the automatic transcription procedure, which was based exclusively on local properties of the speech signal may have selected its symbols less 'systematically' than the human transcribers who may have been biased towards conventional regional characteristics on the basis of subtle cues in the signal. Again, this seems to suggest that we should try and move towards more detailed phonetic transcriptions. As for 3) and 4), we may attempt to classify speakers in more specific classes, hopefully with more homogeneous speech behaviour. In most cases, this is likely to mean a subdivision of the classes used in this study. Recall that we classified our speakers in just 16 predefined geographical regions, and

that we attempted the classification of speakers in just two classes defined by their level of education. The training and use of more specific speaker classes may increase the homogeneity of speech characteristics in these classes, but it would inevitably also introduce a data sparseness problem. Finally, as for 5), we may consider increasing the number of classification features for our algorithm, but we have already argued that it is not obvious how this can be accomplished. Alternatively, we may consider investigating classification techniques that are designed to operate with smaller numbers of features.

Finally, for a real application rather than for a scientific investigation like our study, it will probably be suboptimal to base classifications on a single type of classification features. For the best possible classification, we should give the classifier access to as many and as large a variety of features as possible. This means combining both the lexical and pronunciation features presented here and probably also other features that can be directly retrieved from the speech signal.

Bibliography

Baayen, R.H. (forthcoming). *Analyzing Linguistic Data. A Practical Introduction to Statistics.* Cambridge University Press.

Baayen, R.H., Davidson, D.J., Bates, D.M. (2007). *Mixed Effects Modeling with Crossed Random Effects for Subjects and Items*, submitted manuscript.

Baayen, R.H., Piepenbrock, R., Gulikers, L. (1995). *The CELEX Lexical Database (Release 2).* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, USA.

Bates, D. M. (2005). Fitting Linear Mixed Models in R. In: *R News*, Vol. 5, pp. 27–30.

Bates, D.M., Sarkar, D. (2005). *lme4: Linear Mixed Effects Models using S4 Classes,* R Package Version 0.9975-7.

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. In: *Journal of the Acoustical Society of America*, Vol.113/2, pp. 1001-1024.

Bellegarda, J.R. (2005). Unsupervised, Language-independent Grapheme-to-phoneme Conversion by Latent Analogy. In: *Speech Communication*, Vol. 46/2, pp. 140-152.

Binnenpoorte, D. (2006). *Phonetic Transcriptions of Large Speech Corpora.* PhD Thesis, Radboud University Nijmegen, the Netherlands.

Binnenpoorte, D., Cucchiarini, C. (2003). Phonetic Transcription of Large Speech Corpora: How to Boost Efficiency without Affecting Quality. In: *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 2981-2984.

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

Binnenpoorte, D., Van Bael, C., Os, E. den, Boves, L. (2005). Gender in Everyday Speech and Language: A Corpus-based Study. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp. 2213-2216.

Booij, G. (1999). *The Phonology of Dutch*. Oxford University Press, New York.

Byrd, D. (1994). Relations of Sex and Dialect to Reduction. In: *Speech Communication*, Vol. 15, pp. 39-54.

Catford, J.C. (1974). Phonetic Fieldwork. In: Sebeok, T.A. (Ed.) *Current Trends in Linguistics*, Mouton, The Hague, the Netherlands, Vol. 12, pp. 2489-2505.

CGN - Het Project Corpus Gesproken Nederlands/ The Spoken Dutch Corpus (2005). [http://lands.let.kun.nl/cgn/ehome.htm].

Coussé E., Gillis S. (2006). Regional Bias in the Broad Phonetic Transcriptions of the Spoken Dutch Corpus. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Paris, France, pp. 2080-2083.

CELEX Lexical Database (2005). [http://www.ru.nl/celex/].

Cucchiarini, C. (1993). *Phonetic Transcription: a Methodological and Empirical Study*. PhD Thesis, University of Nijmegen, the Netherlands.

Cucchiarini, C. (1996). Assessing Transcription Agreement: Methodological Aspects. In: *Clinical Linguistics and Phonetics*, Vol. 10/2, pp. 131-155.

Cucchiarini, C., Binnenpoorte, D. (2002). Validation and Improvement of Automatic Phonetic Transcriptions. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, USA, pp. 313-316.

Demuynck, K., Laureys, T., Gillis, S. (2002). Automatic Generation of Phonetic Transcriptions for Large Speech Corpora. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, USA, pp. 333-336.

Demuynck. K., Laureys, T., Wambacq, P., Van Compernolle, D. (2004). Automatic Phonemic Labeling and Segmentation of Spoken Dutch. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 61-64.

Deshpande, M. M. (1994). Ancient Indian Phonetics. In: Asher, R. E., Simpson, J. M. Y. (Eds.) *The Encyclopedia of Language and Linguistics*, Oxford: Pergamon, UK, Vol. 6, pp. 3053-3058.

Elffers, B., Van Bael, C., Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*. Internal Report, Department of Language and Speech, Radboud University Nijmegen, the Netherlands. [http://lands.let.ru.nl/literature/elffers.2005.1.pdf].

Fisher, W., Doddington, G., Goudie-Marshall, K. (1986). The DARPA Speech Recognition Research Database: Specifications and Status. In: *Proceedings of the Defense Advanced Research Projects Agency (DARPA) Speech Recognition Workshop*, SAIC-86/1546, pp. 93-99.

Geumann, A., Oppermann, D., Schaeffler, F. (1997). *The Conventions for Phonetic Transcription and Segmentation of German Used for the Munich Verbmobil Corpus*. Verbmobil Memo 129-96, University of Munich, Germany.

Gillis S., Cucchiarini C., Goddijn S., Pols L. (2001). *Protocol voor Brede Fonetische Transcriptie*. Universiteit Antwerpen (UIA), Antwerp, Belgium. Unpublished Document.

Goddijn, S.M.A., Binnenpoorte, D. (2003). Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. In: *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 1361-1364.

Godfrey, J., Holliman, E., McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, USA, pp. 737-740.

Greenberg, S. (1997). The Switchboard Transcription Project in Research Report #24. In: *1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.

Greenberg, S. (1998). Speaking in Shorthand - A Syllable-centric Perspective for Understanding Pronunciation Variation. In: *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, the Netherlands, pp. 47-56.

Greenberg, S., Hollenback, J., Ellis, D. (1996). Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, pp. S24-27.

Halteren, H. van. (2004). Linguistic Profiling for Author Recognition and Verification. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 200-207.

Halteren, H. van (2007). Author Verification by Linguistic Profiling: An Exploration of the Parameter Space. In: *ACM Transactions on Speech and Language Processing*, Vol. 4/1. [http://portal.acm.org/citation.cfm?id=1217098.1217099].

Harrell F.E. Jr., Lee K.L., Mark D.B. (1996). Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. In: *Statistics in Medicine*, Vol. 15/4, pp.361–387.

Henton, C. (1994). Acoustic Variability in the Vowels of Female and Male Speakers. In: *The Journal of the Acoustical Society of America*, Vol. 94/4, pp. 2387.

Hess, W., Kohler, K.J., Tillman, H.-G. (1995). The Phondat-Verbmobil Speech Corpus. In: *Proceedings of Eurospeech*, 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. In: *International Journal of Speech Technology*, Vol. 4, pp. 127-143.

Heuvel, H. van den, Sanders, E. (2006). Valid Validations: Bare Basics and Proven Procedures. In: *Proceedings of the Workshop "Quality Assurance and Quality Measurement for Language and Speech Resources"*, held at the *International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, (CD-ROM)

Hol, A.R. (2006). Dialectgrenzen in Gelderland. In: Wingens, M.F.M., Demoed, H.B., Scholten, F.W.J. (Eds.) *Gelders Erfgoed, Gelders cultuurhistorisch kwartaalblad*, Vol. 2006/2, pp. 11-13.

Hoste, V., Daelemans, W., Tjong Kim Sang, E., Gillis, S. (2000). Meta-learning for Phonemic Annotation of Corpora. In: *Proceedings of the International Conference on Machine Learning (ICML)*, Stanford University, CA, USA, pp. 375-382.

Howard, S.J., Heselwood, B. (2002). Learning and Teaching Phonetic Transcription for Clinical Purposes. In: *Clinical Linguistics and Phonetics*, Vol. 16, pp. 371-401.

Jande, P.A. (2005). Inducing Decision Tree Pronunciation Variation Models from Annotated Speech Data. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp. 1945-1948.

Johnson, K. (2004). Massive Reduction in Conversational American English. In: Yoneyama, K., Maekawa, K. (Eds.) *Spontaneous Speech: Data and Analysis*. The National Institute for Japanese Language, Tokyo, pp. 29-45.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W.D. (2001). The Effect of Language Model Probability on Pronunciation Variation. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, pp. 801-804.

Kemp, J. A. (1994a). Phonetic Transcription: History. In: Asher, R. E., Simpson, J. M. Y. (Eds.) *The Encyclopedia of Language and Linguistics*, Oxford: Pergamon, UK, Vol. 6, pp. 3040-3051.

Kemp, J. A. (1994b). Phonetics: Precursors of Modern Approaches. In: Asher, R. E., Simpson, J. M. Y. (Eds.) *The Encyclopedia of Language and Linguistics*, Oxford: Pergamon, UK, Vol. 6, pp. 3103-3106.

Kerkhoff, J., Rietveld, T. (1994). Prosody in Niros with Fonpars and Alfeios. In: *Proceedings of the Department of Language and Speech, University of Nijmegen*, Vol. 18, pp. 107-119.

Kessens, J.M. (2002). *Making a difference. On Automatic Transcription and Modeling of Dutch Pronunciation Variation for Automatic Speech Recognition*. PhD Thesis, University of Nijmegen, the Netherlands.

Kessens, J.M., Cucchiarini, C., Strik, H. (2003). A Data-driven Method for Modeling Pronunciation Variation. In: *Speech Communication*, Vol. 40/4, pp. 517-534.

Kessens, J.M., Strik, H. (2004). On Automatic Phonetic Transcription Quality: Lower Word Error Rates Do Not Guarantee Better Transcriptions. In: *Computer Speech and Language*, Vol. 18, pp. 123-141.

Kessens, J.M., Wester, M., Strik, H. (1999). Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation Variation. In: *Speech Communication*, Vol. 29, pp. 193-207.

Keune, K., Ernestus, M., Hout, R. van, Baayen, R.H. (2005). Variation in Dutch: From Written MOGELIJK to Spoken MOK. In: *Corpus Linguistics and Linguistic Theory*, Vol. 1-2, pp. 183-223.

Kipp, A., Wesenick, M.-B., Schiel, F. (1996). Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP),* Philadelphia, USA, pp. 106-109.

Kipp, A., Wesenick, M.-B., Schiel F. (1997). Pronunciation Modelling applied to Automatic Segmentation of Spontaneous Speech. In: *Proceedings of Eurospeech*, Rhodes, Greece, pp. 1023-1026.

Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model of Word-form Recognition and Production*. Technical Report Publication No. 11, Dept. of General Linguistics, University of Helsinki, Finland.

Kuijpers, C., Donselaar, W. van. (1997). The Influence of Rhythmic Context on Schwa Epenthesis and Schwa Deletion in Dutch. In: *Language and Speech*, Vol. 41/1, pp. 87-108.

Labov, W. (1994). *Principles of Linguistic Change*. Blackwell, Cambridge, MA, USA.

Ladefoged, P. (1960). The Value of Phonetic Statements. In: *Language*, Vol. 36, pp. 387-396.

Ladefoged, P. (1993). *A Course in Phonetics - Third Edition.* Harcourt Brace College Pulishers, Forth Worth, TX, USA.

Ladefoged, P. (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Malden, MA: Blackwell Publishing, USA.

Ladefoged, P., Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell Publishers, UK.

Lamel, L.F., Kassel, R.H., Seneff, S. (1986). Speech Database Development: Design and Analysis of the Acoustic-phonetic Corpus. In: *Proceedings of the Defense Advanced Research Projects Agency (DARPA) Speech Recognition Workshop,* SAIC-86/1546, pp.100-109.

Laver, J. (1995). *Principles of Phonetics*. Cambridge University Press, Cambridge, UK.

Ljolje, A., Hirschberg, J., Santen, J.P.H. van (1997). Automatic Speech Segmentation for Concatenative Inventory Selection. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.) *Progress in speech synthesi*s, Springer, New York, USA, pp. 305-311.

Ljolje, A., Riley, M.D. (1991). Automatic Segmentation and Labeling of Speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Canada, pp. S473-S476.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its Design and Evaluation. In: *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, Japan, pp. 7-12.

Mizutani, T., Kagoshima, T. (2005). Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method. In: *IEICE Transactions on Information and Systems*, Vol. E88-D/11, pp. 2565-2572.

Müller, C., Schötz, S. (Eds.) *Speaker Classification*. *Lecture Notes in Computer Science/Artificial Intelligence*, Vol. 4343, Springer, Heidelberg – Berlin - New York. In Press.

Nerbonne, J., Heeringa, W., Hout, E. van den, Kooi, P. van der, Otten, S., Vis, W. van de (1996). Phonetic Distance between Dutch Dialects. In: Durieux, G., Daelemans, W., Gillis, S. (Eds.) *CLIN IV, Papers From the Sixth CLIN Meeting*. University of Antwerp, Center for Dutch Language and Speech, Antwerp, Belgium, pp. 185-202.

Neri, A., Cucchiarini, C., Strik, H. (2006). Selecting Segmental Errors in Non-native Dutch for Optimal Pronunciation Training. In: *International Review of Applied Linguistics*, Vol. 44/4, pp. 357-404.

Neri, A., Cucchiarini, C., Strik, H. (2007). Pronunciation Training in Dutch as a Second Language on the Basis of Automatic Speech Recognition. In: *Stem, Spraak en Taalpathologie*, Vol. 15/11, pp. 159-169.

Neri, A., Cucchiarini, C., Strik, H., Boves, L. (2002). The Pedagogy-Technology Interface in Computer Assisted Pronunciation Training. In: *Computer Assisted Language Learning*, Vol. 15/ 5, pp. 441-467.

Ohala, J. J. (1994). Phonetic Transcription: History. In: Asher, R. E., Simpson, J. M. Y. (Eds.) *The Encyclopedia of Language and Linguistics*, Oxford: Pergamon, UK, Vol. 6, pp. 3040-3053.

Oller, D.K., Eilers, R.E. (1975). Phonetic Expectation and Transcription Validity. In: *Phonetica*, Vol. 31, pp. 288-304.

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

PAROLE lexicon. (2005). [http://ww2.tst.inl.nl].

Pearce, D. (2001). Developing the ETSI Aurora Advanced Distributed Speech Recognition Front-end & What Next? In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Trento, Italy, pp. 131-134.

Pinheiro, J.C., Bates, D.M. (2000) *Mixed-effects models in S and S-PLUS*. Statistics and Computing. Springer, New York.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E. (2006). *Buckeye Corpus of Conversational Speech* (2006; 1st release). [http://www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor), USA.

Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W. (2005). The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. In: *Speech Communication*, Vol. 45/1, pp. 89-95.

Quazza, S., Heuvel, H. van den. (2000). Lexicon Development for Speech and Language Processing. In: Van Eynde, F., Gibbon, D. (Eds.) *Lexicon Development for Speech and Language Processing*. Kluwer Academic Publishers, Dordrecht, pp. 207-233.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann, USA.

R Development Core Team. (2005). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. [http://www.r-project.org/].

Raymond, W.D., Dautricourt, R., Hume, E. (2006). Word-internal /t,d/ Deletion in Spontaneous Speech: Modeling the effects of Extra-linguistic, Lexical, and Phonological Factors. In: *Language Variation and Change*, Vol. 18, pp. 55-97.

Referentiebestand Nederlands (RBN). (2005). [http://ww2.tst.inl.nl].

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje A, McDonough, J., Nock, H., Saraçlar, M., Wooters, C., Zavaliagkos, G. (1999). Stochastic Pronunciation Modelling from Hand-labelled Phonetic Corpora. In: *Speech Communication*, Vol. 29, pp. 209-224.

Saraçlar, M. (2000). *Pronunciation Modeling for Conversational Speech Recognition*. PhD thesis, Johns Hopkins University, Baltimore, MD, USA.

Saraçlar, M., Khundanpur, S. (2004). Pronunciation Change in Conversational Speech and its Implications for Automatic Speech Recognition. In: *Computer Speech and Language*, Vol. 18, pp. 375-395.

Saraçlar, M., Nock, H., Khudanpur, S. (2000). Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models. *Computer Speech and Language*, Vol. 14, pp. 137-160.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In: *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, San Francisco, USA, pp. 607-610.

Shoup, J. E. (1980). Phonological Aspects of Speech Recognition. In: Lea, W.A. (Ed.) *Trends in Speech Recognition*, Prentice-Hall, New York, USA, pp. 125-138.

Shriberg, L.D., Kwiatkowski, J., Hoffman, K. (1984). A Procedure for Phonetic Transcription by Consensus. In: *Journal of Speech and Hearing Research*, Vol. 27, pp. 456-465.

Shriberg, L.D., Lof, L. (1991). Reliability Studies in Broad and Narrow Phonetic Transcription. In: *Clinical Linguistics and Phonetics*, Vol. 5, pp. 225-279.

Straw, W. (1993). Rock Formation: Music, Technology, and Mass Communication. In: *Canadian Journal of Communication*, Vol. 18/4. [http://www.cjc-online.ca/viewarticle.php?id=210].

Strik, H. (Ed.) (1999). Special Issue of Speech Communication on 'Modeling Pronunciation Variation for Automatic Speech Recognition', Vol. 29/2-4.

Strik, H. (2001). Pronunciation Adaptation at the Lexical Level. In: *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) 'Adaptation Methods for Speech Recognition'*, Sophia-Antipolis, France, pp. 123-131.

Strik, H., Cucchiarini, C. (1999). Modeling Pronunciation Variation for ASR: a Survey of the Literature. In: *Special Issue of Speech Communication on 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, Vol. 29/2-4, pp. 225-246.

TIMIT Acoustic-Phonetic Continuous Speech Corpus (1990). National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065, 1990.

Tjalve, M., Huckvale, M. (2005). Pronunciation Variation Modelling using Accent Features. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp.1341-1344.

Van Bael, C., Boves, L., Strik, H., Heuvel, H. van den (2006). Automatic Phonetic Transcription of Large Speech Corpora: a Comparative Study. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh PA, USA, pp. 1085-1088.

Verhoeven, J., De Pauw, G., Kloots, H. (2004). Speech Rate in a Pluricentric Language: A Comparison between Dutch in Belgium and the Netherlands. In: *Language and Speech*, Vol. 47/3, pp. 297-308.

Verstraeten, B., Van de Velde, H. (2001). Socio-geographical Variation of /r/ in Standard Dutch. In: Van de Velde, H., Hout, R. van (Eds.) *r-atic - Sociolinguistic, Phonetic and Phonological Characteristics of /r/.* Etudes & Travaux - ILVP/ULB. No 4. Brussels, pp. 45-61.

Vorstermans, A., Martens, J.P. (1994). Automatic Labeling of Corpora for Speech Synthesis Development. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, pp. 1747-1750.

Wang, L., Zhao, Y., Chu, M., Soong, F., Cao, Z. (2005). Phonetic Transcription Verification with Generalised Posterior Probability. In: *Proceedings of Interspeech*, Lisbon, pp. 1949-1953.

Wells, J.C. (1997). SAMPA Computer Readable Phonetic Alphabet. In: Gibbon, D., Moore, R., Winski, R. (Eds.) *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.

Wells, J.C. (2000). *Longman Pronunciation Dictionary*. Second edition. Harlow: Pearson Education Limited, UK.

Wester, M. (2002). *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*. PhD Thesis, University of Nijmegen, the Netherlands.

Wester, M. (2003). Pronunciation Modeling for ASR - Knowledge-based and Data-derived Methods. In: *Computer Speech and Language*, Vol. 17/1, pp. 69-85.

Witten, I.H., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Morgan Kaufmann, San Francisco, USA.

Yang, Q., Martens, J.-P. (2000). Data-driven Lexical Modelling of Pronunciation Variations for ASR. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, pp. 417-420.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P. (2001). *The HTK Book (for HTK version 3.2)*, Cambridge University Engineering Department, UK.

Summary

This section provides a brief summary of the six chapters in this thesis.

Chapter 1 – Introduction

Phonetic transcriptions represent the pronunciation of words as strings of characters from specifically designed symbol sets. More elaborate symbol sets allow for the representation of more phonetic detail. The degree of phonetic detail in phonetic transcriptions is an important factor in determining which purposes the transcriptions can and cannot be used for.

This dissertation reports work on and with so-called *broad phonetic transcriptions*. Broad phonetic transcriptions are capable of describing only the most general pronunciation variation in speech, but they have nonetheless proven useful for phonetics, phonology, sociolinguistics, language pedagogy, lexicography, for the study of speech and language disorders and for speech therapy. Moreover, in the last three decades, broad phonetic transcriptions have also been used in computer-driven speech applications such as computer assisted pronunciation training, automatic speech recognition and text-to-speech synthesis.

Broad phonetic transcriptions can be generated manually, automatically and semi-automatically. Manual transcriptions are usually preferred over automatic and semi-automatic transcriptions, because they are entirely based on the auditory perception and the expertise of human transcribers. However, the fully manual transcription of speech has proven to be time-consuming and expensive. Therefore, large speech corpora are often transcribed by means of a semi-automatic transcription procedure in which human transcribers verify and correct an automatically generated hypothetical transcription, instead of transcribing the speech signal from scratch. Such check-and-correct procedures have proven to be much quicker and cheaper than fully manual transcription, but they have also shown to imply the risk of generating transcriptions that are biased towards the hypothetical transcription they are based upon. In addition, such procedures still require a considerable amount of time and effort because of the involvement of human transcribers.

Phonetic transcriptions are customarily evaluated in terms of their *validity*, i.e. the accuracy with which they describe the auditory perception of speech as strings of phonetic symbols. Experience shows that, in particular in large transcription projects, phonetic transcriptions are usually validated in terms of their resemblance to one or more handcrafted reference transcriptions, irrespective of the procedure by means of which the transcriptions were generated and irrespective of the application(s) the transcriptions will be used for. Since phonetic transcriptions are often generated to serve one or a few particular purposes or applications which are not necessarily all linguistically oriented, it seems sensible to validate phonetic transcriptions in terms of measures that are directly relevant for the purpose(s) they will be used for, instead of in terms of their resemblance to a fixed reference transcription.

Chapter 2 - Validation of Phonetic Transcriptions in the Context of Automatic Speech Recognition

Phonetic transcriptions are nowadays widely used for the development of automatic speech recognition systems. Developers of such systems usually use phonetic transcriptions of existing speech corpora of which the validity was previously assessed in terms of their resemblance to a purpose-independent handcrafted reference transcription. Although some applications may fare best with transcriptions that closely resemble such a general reference transcription, it is not clear whether this also holds for automatic speech recognition. In the experiment described in Chapter 2, we verified whether it is safe to validate phonetic transcriptions in terms of their similarity to a purpose-independent reference transcription if the transcriptions are to be used for the development of automatic speech recognition systems.

To this end, we evaluated two types of transcriptions (a canonical representation and a semi-automatic (viz. manually verified) phonetic transcription) of well-prepared and spontaneous speech in terms of their resemblance to a handcrafted reference transcription on the one hand, and in terms of their suitability for ASR development on the other hand. Then we compared whether the two evaluations attributed the same validity rating to both types of transcriptions. Whereas the manually verified phonetic transcriptions resembled the reference transcription much closer than the canonical representations, the use of both transcription types yielded similar recognition results. The difference between the outcomes of the two evaluations has two implications.

First, this result implies that whenever possible, the validation of phonetic transcriptions should be carried out in terms of the quality measure of the application the transcriptions will be used for. In addition, it implies that in spite of the high costs and the time required to generate them, manually verified phonetic transcriptions are not necessarily preferable for the development of automatic speech recognition systems; we obtained similar recognition results with a recognition system that was developed with a much cheaper canonical representation of speech.

Chapter 3 – Automatic Phonetic Transcription of Large Speech Corpora

The employment of human transcribers for the manual transcription of large speech corpora is (usually too) time-consuming and expensive. A common workaround is to employ human transcribers for the verification of automatically generated transcriptions instead of for fully manual transcription from scratch. The Spoken Dutch Corpus is one such corpus that was partially provided with a manually verified phonetic transcription. Experience indeed showed that the verification of a canonical transcription was considerably quicker than manual transcription from scratch. Nevertheless, it still took a considerable amount of time to manually verify the automatically generated transcriptions of a 1-million-word subset of the corpus. Therefore, we investigated whether manually verified transcriptions such as the transcriptions of the Spoken Dutch Corpus can be approximated by means of a fully automatic transcription procedure. Should this be the case, it would imply a considerable time gain and cost reduction for future transcription projects.

We used several automatic transcription procedures to generate transcriptions of a small sample of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The transcriptions were compared with the corresponding manually verified phonetic transcriptions of the Spoken Dutch Corpus. We found an automatic transcription procedure that by means of decision trees and a small seed corpus of target transcriptions proved capable of approximating the quality of the manually verified phonetic transcriptions. Considering that most of the remaining discrepancies between the automatic transcriptions and the manually verified phonetic transcriptions can probably be attributed to uncertainties that are typical of human transcription behaviour, our study suggests that it is sufficient to verify the phonetic transcription of only a small portion of a corpus by hand in order to automatically generate similar transcriptions for the remainder of the corpus by means of machine learning algorithms such as decision trees.

Chapter 4 – Segment Deletion in Spontaneous Speech: A Corpus Study using Mixed Effects Models with Crossed Random Effects

Over the years, large annotated speech corpora such as Switchboard and the Buckeye Corpus of Conversational Speech have proven useful for -among other things- creating an inventory of testified speech processes in everyday conversational English, studying the frequencies of these processes and investigating how these processes are influenced by various linguistic and sociolinguistic factors. Because most phonetically transcribed speech corpora comprise (American) English, most corpus studies on pronunciation variation were conducted on English. The recent release of the richly annotated 9-million-word Spoken Dutch Corpus now offers new opportunities for studying pronunciation variation in a language other than English, and for testing whether findings for English generalise to another language.

The first aim of our study was to establish the frequencies of segment deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the linguistic and sociolinguistic factors reflected in the annotations, word segmentations and metadata of the CGN. We defined segment deletion as the deletion of phones and syllables that can be inferred from the symbolic alignment of canonical and manually verified phonetic transcriptions from the so-called *core corpus* of the Spoken Dutch Corpus. We found average phone and syllable deletion rates of 7.57% and 5.46% respectively. 22.76% of the content words and 18.10% of the function words had at least one phone missing, and 6.66% of the content words and 7.10% of the function words had at least one syllable missing. Yet, the total number of phone deletions in function words was larger than in content words. This can be explained by the larger number of function words that had multiple phones missing. These results are in line with findings reported for American English. Phone and syllable deletions are just as well common in everyday conversational Dutch as they are in English.

An ancillary goal of our study was to explore the new opportunities for analysing complex corpus data offered by a recently developed statistical modelling technique: the possibility to

model the effects of random factors as crossed instead of nested with generalised linear mixed effects models (GLMMs). Mixed effects models are interesting for linguistic corpus studies because they allow for the inclusion of factors with repeatable levels (e.g. word class) and randomly sampled levels (e.g. speaker) in the same model, because they can cope with missing data and with complex factorial designs, and because they can do all this in a computationally efficient way. Until recently, however, factors with randomly sampled levels could only be modelled with nested designs. This imposed serious limitations on the use of mixed-effects models for linguistic studies, because it could increase the risk of type I errors, i.e. erroneously considering an effect significant. The recent possibility to model random effects as crossed instead of nested alleviates this problem. The mixed effects models we fitted for phone and syllable deletion had several effects in common, which implies that both types of deletion are to a large extent influenced by the same factors. The strongest factors across both models were lexical stress, word duration and the segmental context of the syllable onset of the following word.

Because we could include random-effects factors such as speaker, word and syllable identity as crossed instead of nested, we were able to assess in a methodologically sound way the relative effect of every linguistic and sociolinguistic factor in the annotations, word segmentations and metadata of the Spoken Dutch Corpus over and above the random variation that came with the speakers, words and syllables we sampled. In our study, it was not only interesting to analyse which factors were significant in the models, but it was equally interesting to see that the (potential) effects of factors which were previously reported to influence segment deletion were 'covered' by other factors. For example, mutual information (word predictability) which was previously reported to influence phone deletions did not appear in our final model definitions, and word frequency was only significant in the phone deletion model. This implies that in our model definitions other factors showed a stronger effect on the deletion of phones and syllables. Actually, word frequency was part of the syllable model definition until we included 'word identity' as random effects factor. In both models, the effects of mutual information were probably covered by word frequency. Such knowledge is unlikely to be gained in controlled experiments on selected data sets aimed at studying the effects of one or a few factors at a time, but it can be of interest for pronunciation variation modelling in everyday conversational speech.

Chapter 5 – Speaker Classification by means of Orthographic and Broad Phonetic Transcriptions of Speech

In Chapter 5, we attempted automatic speaker classification on the basis of orthographic and automatically generated phonetic transcriptions. Our approach differed from conventional speaker classification approaches in that our classification algorithm did not classify speakers on the basis of direct acoustic measurements of the speech signal, but on the basis of written representations of the signal instead.

The classification algorithm first analysed orthographic and automatic broad phonetic transcriptions of speech from speakers whose gender, age, regional background and level of education were represented in the metadata of the Spoken Dutch Corpus. From these analyses, the algorithm identified regularities in the use of words and pronunciation processes of speakers from specific speaker classes (e.g. male speakers, highly educated speakers, speakers from Groningen). Per speaker class, these regularities were organised into two separate 'classification profiles': one lexical profile representing class-specific use of words, and a pronunciation profile representing class-specific pronunciation processes. These classification profiles were subsequently used to accept or reject unknown speakers as members of various speaker classes on the basis of the orthographic and broad phonetic transcriptions of their speech.

The classification algorithm proved able to retrieve and use characteristic features from the orthographic transcription. However, the algorithm's performance with the lexical profiles was probably not good enough for operational speaker classification: in general, we found equal error rates between 20% and 40%. When the classification algorithm had access only to the pronunciation features it extracted from the automatic broad phonetic transcription, it was hardly ever able to classify speakers effectively. We have given several possible explanations for this, and we have argued how these suggestions can be investigated in future research.

Chapter 6 – General Discussion

The studies in this dissertation were conducted with three aims in mind. These aims concerned the validation, the automatic generation and the use of broad phonetic transcriptions.

In Chapter 2, we reconsidered the general feasibility of procedures that validate phonetic transcriptions by means of a comparison with a purpose-independent human-made reference transcription. The experiments showed that canonical transcriptions can serve the purpose of developing basic automatic speech recognition systems equally well as more expensive manually verified phonetic transcriptions that resembled a handcrafted reference transcription better. This finding can probably be explained by a specific feature of statistical pattern recognition: using canonical representations both for training and testing speech recognition systems results in a smaller training-test mismatch than using manual transcriptions for training and canonical representations for testing. On the basis of this finding, we have argued that developers of automatic speech recognition systems need to invest in the manual verification of phonetic transcriptions only for specific purposes, e.g. for in-depth analyses of what caused recognition errors, or for modelling pronunciation variation. A second, more important implication of our results is that phonetic transcriptions should preferably be validated in the context of the application they will serve, because a higher resemblance to a purpose-independent reference transcription proved no guarantee for a transcription to be better suited for the development of automatic speech recognition systems.

In Chapter 3, we investigated to what extent one can approximate expensive manually verified phonetic transcriptions with a fully automatic and therefore quicker and cheaper transcription procedure. We found a procedure that by means of decision trees and a small

seed corpus of target transcriptions proved capable of approximating the manually verified phonetic transcriptions of the Spoken Dutch Corpus. This finding may be of use of for future corpus designers, because the use of machine learning techniques such as decision trees may offer a quick and cheap alternative to employing human transcribers for the manual verification of automatically generated transcriptions of large speech corpora. Our study suggests that it is sufficient to verify the phonetic transcription of only a small portion of a corpus by hand in order to automatically generate similar transcriptions for the remainder of the corpus by means of decision trees. Even though our procedure requires the availability of a small seed corpus of target transcriptions, the generation of such a small set of transcriptions is almost always (much) cheaper than providing such transcriptions for a full corpus.

In Chapter 4, we studied the statistics of phone and syllable deletions in spontaneous Dutch on the basis of manually verified phonetic transcriptions from the Spoken Dutch Corpus. In addition, statistical analyses of these transcriptions in the context of the other annotations and metadata of the Spoken Dutch Corpus showed to which extent phone and syllable deletions are influenced by the interplay of various (socio)linguistic factors. The statistical analyses were made possible by a recent development in computational statistics: the possibility to model random effects in mixed effects models in a principled way as crossed instead of nested. The availability of large and richly annotated speech corpora on the one hand, and the new possibility to model random effects as crossed in mixed effects models on the other, now makes it possible to study linguistic phenomena such as segment deletion as a function of the interplay of many factors instead of in controlled experimental environments designed for studying the effects of one or a few factors at a time. Since mixed effects models with crossed random effects can be fitted to all kinds of data with random and fixed effects, this technique offers new promising opportunities for many kinds of linguistic studies on large and richly annotated speech corpora.

In Chapter 5, we investigated whether a classification algorithm originally designed for authorship verification can also be used to classify speakers in terms of their gender, age, regional background and level of education by analysing orthographic and automatically generated broad phonetic transcriptions of their speech. The classification algorithm was able to retrieve and use characteristic features from the orthographic transcriptions. Despite some encouraging results, however, the algorithm's performance was not good enough for operational speaker classification: in general, we found equal error rates between 20% and 40%. When the classification algorithm had access only to the automatic phonetic transcription, it was hardly ever able to classify speakers effectively. We hypothesised that this may be explained by 1) the absence of distinguishing pronunciation features at the broad phonetic level, 2) the failure of the automatic phonetic transcription procedure to capture distinguishing pronunciation features, 3) a mismatch between our speaker classes and groups of speakers that possibly show distinguishing speech features, 4) the heterogeneity of our speaker classes (either because they are inherently heterogeneous or because the speakers were not representative of their classes), and 5) the limitations of our algorithm for classification with a small number of classification features. We defined suggestions for future research to verify these hypotheses.

Samenvatting (summary in Dutch)

Deze sectie geeft een korte samenvatting van de zes hoofdstukken in dit proefschrift.

Hoofdstuk 1 – Inleiding

Fonetische transcripties representeren de uitspraak van woorden door middel van symbolen uit een speciaal ontworpen alfabet. Hoe uitgebreider het alfabet, hoe meer fonetisch detail ermee gerepresenteerd kan worden. De mate van detail in fonetische transcripties bepaalt in grote mate voor welke doelen de transcripties wel of niet gebruikt kunnen worden.

In dit proefschrift zijn vier onderzoeken opgenomen over en met zogenaamde *brede fonetische transcripties*. Hoewel brede fonetische transcripties slechts de meest algemene kenmerken van spraak kunnen representeren, hebben ze hun nut al bewezen voor fonetische, fonologische en sociolinguïstische studies, voor taalonderwijs, lexicografie, voor de studie van spraak- en taalstoornissen en voor logopedie. Sinds de jaren tachtig van de vorige eeuw worden brede fonetische transcripties bovendien vaak gebruikt in toepassingen van spraaktechnologie zoals computergebaseerde uitspraaktraining, automatische spraakherkenning en spraaksynthese.

Brede fonetische transcripties kunnen handmatig, automatisch of semi-automatisch gegenereerd worden. Handmatig gegenereerde fonetische transcripties genieten meestal de voorkeur boven automatisch gegenereerde transcripties omdat zij volledig gebaseerd zijn op de auditieve perceptie en de kunde van menselijke transcribenten. Helaas is de manuele transcriptie van spraak tijdrovend en daarom ook duur. Daarom worden grote spraakcorpora vaak getranscribeerd door middel van een semi-automatische procedure waarin menselijke transcribenten een automatische fonetische transcriptie verifiëren en waar nodig corrigeren. Hoewel zulke semi-automatische procedures sneller en goedkoper zijn dan volledig handmatige procedures, kost het menselijke transcribenten vaak nog veel tijd om de fonetische transcripties van een grote verzameling spraakmateriaal te verifiëren. Bovendien houden zulke procedures het risico in dat de uiteindelijke transcripties onterecht veel lijken op de automatisch gegenereerde transcripties die geverifieerd werden.

Fonetische transcripties worden meestal geëvalueerd in termen van hun *validiteit* of de accuraatheid waarmee ze de auditieve perceptie van spraak beschrijven met een rij fonetische symbolen. In de praktijk wordt de validiteit van fonetische transcripties meestal bepaald in termen van hun gelijkenis met één of meer handgemaakte referentietranscripties, ongeacht de procedure waarmee de fonetische transcripties gemaakt werden, en ongeacht de toepassing(en) waarvoor de transcripties gebruikt zullen worden. Aangezien fonetische transcripties vaak gemaakt worden voor één of een aantal specifieke doelen of toepassingen die niet noodzakelijk linguïstisch van aard zijn, lijkt het logisch(er) om fonetische transcripties te valideren in de context van deze doelen en toepassingen in plaats van in de context van hun gelijkenis met een vaste referentietranscriptie.

Hoofdstuk 2 – Validatie van Fonetische Transcripties voor Automatische Spraakherkenning

Fonetische transcripties worden tegenwoordig vaak gebruikt voor de ontwikkeling van automatische spraakherkenners. Ontwerpers van zulke systemen gebruiken meestal fonetische transcripties uit bestaande spraakcorpora, transcripties dus waarvan de validiteit eerder al bepaald werd op basis van hun gelijkenis met een handgemaakte referentietranscriptie. Hoewel sommige toepassingen misschien beter werken met transcripties die erg lijken op een handgemaakte referentietranscriptie, is het niet duidelijk of dit ook geldt voor automatische spraakherkenning. In het tweede hoofdstuk van dit proefschrift hebben we daarom geverifieerd of het verstandig is om fonetische transcripties te valideren in termen van hun gelijkenis met een algemene referentietranscriptie als de transcripties gebruikt zullen worden voor de ontwikkeling van een automatische spraakherkenner.

We evalueerden twee soorten fonetische transcripties (een canonieke representatie en een manueel geverifieerde transcriptie) van voorgelezen en spontane spraak in termen van hun gelijkenis met een handgemaakte referentietranscriptie enerzijds, en in termen van hun geschiktheid voor de ontwikkeling van een automatische spraakherkenner anderzijds. Vervolgens vergeleken we de uitkomst van de evaluaties. Hoewel de handmatig geverifieerde transcripties meer gelijkenissen vertoonden met de referentietranscriptie dan de canonieke representaties (en dus een hogere validiteit toebedeeld kregen), werden op basis van beide types transcripties vergelijkbare herkenresultaten geboekt. Dit resultaat heeft twee implicaties.

In de eerste plaats impliceert dit resultaat dat fonetische transcripties bij voorkeur gevalideerd moeten worden in de context van het kwaliteitscriterium van de toepassing waarvoor zij gebruikt zullen worden. Verder impliceert dit resultaat dat ongeacht de hoge kosten en de tijd die vereist zijn om manueel geverifieerde fonetische transcripties te maken, zulke transcripties niet noodzakelijk beter geschikt zijn voor de ontwikkeling van automatische spraakherkenners; we behaalden vergelijkbare herkenresultaten met een spraakherkenner die ontwikkeld was met een veel goedkopere canonieke representatie.

Hoofdstuk 3 – Automatische Fonetische Transcriptie van Grote Spraakcorpora

Het is meestal te tijdrovend en te duur om menselijke transcribenten in te huren voor de handmatige transcriptie van grote spraakcorpora. In plaats daarvan worden transcribenten vaak ingehuurd voor de verificatie van automatisch gegenereerde voorbeeldtranscripties. Het Corpus Gesproken Nederlands (CGN) is een recent corpus dat gedeeltelijk voorzien werd van een dergelijke manueel geverifieerde fonetische transcriptie. Al snel bleek echter dat ook een verificatieprocedure nog behoorlijk veel tijd in beslag kan nemen. Daarom onderzochten we met het oog op toekomstige transcriptieprojecten of het mogelijk is om de kwaliteit van

manueel geverifieerde transcripties zoals die van het CGN te benaderen met behulp van een volledig automatische en dus ook snellere en goedkopere transcriptieprocedure.

We gebruikten verschillende automatische transcriptieprocedures om een brede fonetische transcriptie te genereren van een kleine hoeveelheid voorgelezen en spontane spraak (voorgelezen verhalen en telefoongesprekken) uit het CGN. Vervolgens werden deze transcripties vergeleken met de manueel geverifieerde fonetische transcripties van het CGN. We vonden een automatische transcriptieprocedure die met behulp van beslissingsbomen en een kleine hoeveelheid voorbeeldtranscripties de kwaliteit van de manueel geverifieerde fonetische transcripties van het CGN kan benaderen. De resterende verschillen tussen de automatisch gegenereerde transcripties en de manueel geverifieerde fonetische transcripties zijn grotendeels vergelijkbaar met de verschillen die vaak ook tussen de transcripties van menselijke transcribenten gevonden worden. Dit suggereert dat men kan volstaan met de handmatige verificatie van een kleine hoeveelheid spraakmateriaal van een corpus om vervolgens de rest van het corpus automatisch te voorzien van een gelijkaardige transcriptie.

Hoofdstuk 4 – Segmentdeletie in Spontane Spraak: Een Corpusonderzoek met 'Mixed Effects Models' met 'Crossed Random Effects'

Door de jaren heen hebben grote geannoteerde spraakcorpora zoals Switchboard en het Buckeye Corpus of Conversational Speech hun nut bewezen voor onder andere het inventariseren van uitspraakprocessen en voor het bestuderen van de frequenties van deze processen en de manier waarop deze processen beïnvloed worden door verschillende linguïstische en sociolinguïstische factoren. Omdat de meeste fonetisch getranscribeerde spraakcorpora Engelse spraak bevatten werden de meeste corpusstudies naar uitspraakvariatie uitgevoerd op het Engels. De recente oplevering van het uitvoerig geannoteerde Corpus Gesproken Nederlands biedt nu de mogelijkheid om uitspraakvariatie te bestuderen in een andere taal dan het Engels, en om te testen of onze kennis over uitspraakvariatie in het Engels ook geldt voor het Nederlands.

Het eerste doel van ons onderzoek was vast te stellen hoe vaak klanken en lettergrepen *niet* uitgesproken worden in spontane Nederlandse spraak, en in welke mate zulke 'deleties' (ten opzichte van de standaarduitspraak van woorden) beïnvloed worden door linguïstische en sociolinguïstische factoren die kunnen afgeleid worden uit de annotaties, woordsegmentaties en metadata van het CGN. We lokaliseerden segmentdeleties door middel van een symbolische oplijning van canonieke en manueel geverifieerde fonetische transcripties van woorden in het CGN. We vonden dat 7.57% van alle klanken en 5.46% van alle lettergrepen in de standaarduitspraak van de woorden niet uitgesproken werden. In 22.76% van de inhoudswoorden en in 18.10% van de functiewoorden ontbrak minstens één klank, en 6.66% van de inhoudswoorden en 7.10% van de functiewoorden hadden minstens één ontbrekende lettergreep. Toch was het totale aantal klankdeleties in functiewoorden groter dan in inhoudswoorden; in functiewoorden komt het veel vaker voor dat meer dan één klank niet

gerealiseerd wordt. Deze resultaten liggen op dezelfde lijn als resultaten die vroeger al gerapporteerd werden voor het Engels. Het niet uitspreken van klanken en lettergrepen in spontane spraak is dus blijkbaar net zo gebruikelijk in het Nederlands als in het Engels.

Een bijkomend doel van ons onderzoek was gericht op het onderzoeken van de bruikbaarheid van een nieuwe statistische techniek voor het analyseren van complexe corpusgegevens. Sinds kort is het mogelijk om in 'generalised linear mixed effects models' (GLMMs) 'random factors' in een 'crossed' en dus niet langer enkel in een 'nested' design te modelleren. GLMMs zijn intrinsiek interessant voor het uitvoeren van taalkundig onderzoek op grote hoeveelheden corpusmateriaal omdat deze modellen het mogelijk maken om factoren met herhaalbare waarden (zoals woordklasse) en factoren met willekeurig gesamplede waarden (zoals spreker) te modelleren in één model, omdat ze met ontbrekende gegevens kunnen omgaan, en omdat ze dit alles kunnen op een computationeel efficiënte manier. Tot voor kort echter konden verschillende random-effects factoren enkel in een nested design gemodelleerd worden. Dit beperkte de bruikbaarheid van GLMMs voor taalkundige studies omdat het de kans op type I fouten (d.w.z. de invloed van een factor op een taalkundig fenomeen onterecht als significant beschouwen) vergrootte. De recente mogelijkheid om random factors als crossed in plaats van nested te modelleren lost dit probleem grotendeels op. De GLMMs waarmee we klank- en lettergreepdeleties modelleerden hadden verschillende factoren gemeen. Dit wil zeggen dat volgens onze modellen klank- en lettergreepdeleties in spontane Nederlandse spraak voor een groot deel beïnvloed worden door dezelfde factoren. De sterkste factoren die een invloed hadden op zowel klank- als syllabedeleties waren lexicale klemtoon, woordduur en de eerste klank (consonant of vocaal) van het volgende woord.

Omdat we random-effects factoren zoals spreker, woord en lettergreep als crossed in plaats van nested konden modelleren, konden we op een methodologisch verantwoorde manier de relatieve effecten van elke linguïstische en sociolinguïstische factor bepalen, ongeacht de willekeurige variatie in de deleties die resulteerde uit de grote verscheidenheid aan sprekers, woorden en lettergrepen die we onderzochten. Het bleek niet alleen interessant om te analyseren welke factoren een significante invloed uitoefenden op het niet uitspreken van klanken. Het was net zo interessant om te merken dat de (potentiële) effecten van bepaalde factoren overschaduwd werden door de effecten van andere factoren. Mutual information bijvoorbeeld (de voorspelbaarheid van woorden gegeven de omliggende woorden) werd in het verleden vaak als invloedrijke factor beschouwd voor het niet uitspreken van klanken en syllabes. Deze factor zat echter niet in onze uiteindelijke modellen voor klank- en syllabedeletie. Dit houdt in dat in onze modeldefinities de invloeden van andere factoren (m.n. woordfrequentie) op klankdeletie harder doorwogen dan de invloed van mutual information. Zulke kennis is moeilijk te verkrijgen op basis van gecontroleerde experimenten op zorgvuldig geselecteerde spraakbestanden, maar kan wel interessant zijn voor uitspraakmodellering van alledaagse spontane spraak.

Hoofdstuk 5 – Sprekerclassificatie op basis van Orthografische en Brede Fonetische Transcripties van Spraak

Hoofdstuk 5 beschrijft een experiment met automatische sprekerclassificatie op basis van orthografische en automatisch gegenereerde brede fonetische transcripties. Onze aanpak verschilde van de aanpak in traditionele sprekerclassificatieprocedures omdat ons algoritme de sprekers niet classificeerde op basis van directe akoestische analyses van het spraaksignaal, maar op basis van symbolische representaties van dat signaal.

Het algoritme analyseerde eerst orthografische en automatisch gegenereerde fonetische transcripties van spraak van sprekers waarvan het geslacht, de leeftijd, de regionale achtergrond en het opleidingsniveau bekend waren in de metadata van het Corpus Gesproken Nederlands. Uit deze analyses extraheerde het algoritme kenmerkend woordgebruik en specifieke uitspraakprocessen voor verschillende sprekerklassen (bijvoorbeeld mannelijke sprekers, hoog opgeleide sprekers, sprekers uit Groningen). Deze kenmerken werden per sprekerklasse opgeslagen in twee 'classificatieprofielen': één profiel dat kenmerkend woordgebruik representeerde, en één profiel dat kenmerkende uitspraakprocessen voorstelde. Deze classificatieprofielen werden vervolgens gebruikt om nieuwe sprekers op basis van een analyse van een orthografische en een fonetische transcriptie van hun spraak in te delen in de verschillende sprekerklassen.

Het classificatiealgoritme kon kenmerkende eigenschappen extraheren uit de orthografische transcripties. De accuraatheid van het algoritme was echter niet hoog genoeg om het te kunnen gebruiken voor praktische sprekerclassificatie. Het algoritme had ook grote moeite om sprekers te classificeren op basis van de uitspraakkenmerken die uit de automatische fonetische transcriptie geëxtraheerd werden. We hebben hier verschillende verklaringen voor gegeven en we hebben suggesties gedaan voor verder onderzoek naar deze verklaringen.

Hoofdstuk 6 – Algemene Discussie

De studies in dit proefschrift werden uitgevoerd met drie doelen voor ogen. Deze doelen hadden betrekking op de validatie, de automatische productie en het gebruik van brede fonetische transcripties.

In hoofdstuk 2 verifieerden we of het verstandig is om fonetische transcripties te valideren op basis van hun gelijkenis met een handgemaakte referentietranscriptie. De experimenten toonden aan dat canonieke transcripties net zo geschikt zijn voor de ontwikkeling van een standaard automatische spraakherkenner als duurdere manueel geverifieerde transcripties die meer gelijkenis vertonen met een handgemaakte referentietranscriptie. We concludeerden dat ontwikkelaars van spraakherkenners beter enkel nog voor specifieke doeleinden investeren in de manuele verificatie van fonetische transcripties. Hierbij kan gedacht worden aan grondige analyses om te bepalen hoe bepaalde herkenfouten tot stand komen, of aan fonetische

transcripties die als hulpmiddel moeten dienen bij de modellering van uitspraakvariatie. Bovendien, en belangrijker nog, impliceerde dit resultaat dat fonetische transcripties het beste gevalideerd kunnen worden in de context van de toepassing waarvoor ze gebruikt zullen worden. Een transcriptie die meer leek op een handgemaakte referentietranscriptie bleek immers niet noodzakelijk beter geschikt voor de ontwikkeling van een automatische spraakherkenner.

In Hoofdstuk 3 onderzochten we in welke mate de kwaliteit van dure manueel geverifieerde fonetische transcripties benaderd kan worden met een volledig automatische en daardoor snellere en goedkopere transcriptieprocedure. We vonden een eenvoudige procedure die met behulp van beslissingsbomen en een kleine hoeveelheid nauwkeurige transcripties de kwaliteit van de manueel geverifieerde fonetische transcripties van het Corpus Gesproken Nederlands kon benaderen. Deze bevinding kan belangrijk zijn voor de transcriptie van toekomstige spraakcorpora, omdat het gebruik van zelflerende systemen zoals beslissingsbomen een snel en goedkoop alternatief kan bieden voor het inhuren van menselijke transcribenten voor de handmatige verificatie van voorbeeldtranscripties. Ons onderzoek suggereert dat het voldoende is om een nauwkeurige fonetische transcriptie van een klein deel van een corpus te maken, om vervolgens volledig automatisch een gelijkaardige transcriptie te maken van de rest van het corpus. Dit kan leiden tot een behoorlijke tijdswinst en een verlaging van de transcriptiekosten.

In Hoofdstuk 4 bestudeerden we de deletie van klanken en lettergrepen in spontane Nederlandse spraak op basis van analyses van manueel geverifieerde fonetische transcripties van het CGN. Deze transcripties bleken een goede bron van informatie te zijn om de frequentie van foon- en lettergreepdeleties te bestuderen, en bovendien konden we deze transcripties ook gebruiken om te modelleren tot op welke hoogte zulke deleties door het samenspel van verschillende (socio)linguïstische factoren bepaald worden. We konden onze analyses uitvoeren dankzij een recente ontwikkeling in de computationele statistiek: de mogelijkheid om ‘random effects’ in ‘mixed effects’ modellen als ‘crossed’ in plaats van ‘nested’ te modelleren. De beschikbaarheid van grote en uitgebreid geannoteerde spraakcorpora enerzijds en de nieuwe mogelijkheid om willekeurige effecten als ‘crossed’ te modeleren in ‘mixed effects models’ anderzijds zorgen ervoor dat het nu mogelijk is om taalkundige verschijnselen zoals segmentdeletie te bestuderen in functie van het samenspel van vele factoren in plaats van in gecontroleerde experimentele omgevingen die opgezet zijn om het effect van één of een klein aantal factoren te bestuderen. Aangezien ‘mixed effects models’ gebruikt kunnen worden voor het modelleren van ‘random’ en ‘fixed effects’ biedt deze techniek nieuwe mogelijkheden voor taalkundige studies op grote en uitgebreid geannoteerde spraakcorpora.

In Hoofdstuk 5 onderzochten we of een automatisch classificatiealgoritme sprekers uit het Corpus Gesproken Nederlands kon classificeren volgens hun geslacht, leeftijd, regionale achtergrond en opleidingsniveau op basis van een analyse van hun uitspraak en woordgebruik. Onze aanpak verschilde van de aanpak in traditionele procedures omdat ons algoritme de sprekers niet classificeerde op basis van directe akoestische analyses van het spraaksignaal, maar op basis van symbolische representaties van dat signaal. Het algoritme bleek in staat om karakteristieke lexicale kenmerken uit de orthografische transcripties te extraheren. Ondanks

deze bemoedigende resultaten echter maakte het algoritme op basis van deze informatie teveel fouten om het in de praktijk voor sprekerclassificatie te kunnen gebruiken. Het algoritme had zelfs nog meer moeite om sprekers te classificeren op basis van uitspraakkenmerken die uit de automatisch gegenereerde brede fonetische transcripties geëxtraheerd werden. We argumenteerden dat dit mogelijk verklaard kan worden door 1) de afwezigheid van onderscheidende uitspraakprocessen op het brede fonetische niveau, 2) tekortkomingen in onze automatische transcriptieprocedure om eventuele onderscheidende uitspraakprocessen te representeren, 3) een discrepantie tussen de door ons gedefinieerde sprekerklassen en de groepen sprekers die in de realiteit onderscheidende spraakkenmerken vertonen, 4) de heterogeniteit van onze sprekerklassen (omdat de klassen in de praktijk inherent heterogeen zijn of omdat de sprekers niet representatief waren voor hun klassen), en 5) de beperkingen van ons algoritme voor classificatie met een beperkt aantal classificatiekenmerken. We definieerden enkele vervolgstudies om deze verklaringen verder te onderzoeken.

List of Publications

This thesis comprises the following publications:

Van Bael, C., Heuvel, H. van den, Strik, H. (in press). Validation of Phonetic Transcriptions in the Context of Automatic Speech Recognition. In: *Language Resources and Evaluation*.

Van Bael, C., Boves, L., Heuvel, H. van den, Strik, H. (2007). Automatic Phonetic Transcription of Large Speech Corpora. In: *Computer Speech and Language*, Vol. 21, pp. 652-668.

Van Bael, C., Baayen, R.H., Strik, H. (to be submitted). Segment Deletion in Spontaneous Speech: A Corpus Study Using Mixed Effects Models with Crossed Random Effects.

Van Bael, C., Halteren, H. van (in press). Speaker Classification by means of Orthographic and Broad Phonetic Transcriptions of Speech. In: Müller, C., Schötz, S. (Eds.) *Speaker Classification*. *Lecture Notes in Computer Science/Artificial Intelligence*, Vol. 4343, Springer, Heidelberg – Berlin - New York.

Other publications not included in this thesis:

Van Bael, C., Baayen, R.H., Strik, H. (2007). Segment Deletion in Spontaneous Speech: A Corpus Study Using Mixed Effects Models with Crossed Random Effects. In: *Proceedings of Interspeech – Eurospeech*, Antwerp, Belgium.

Van Bael, C., Boves, L., Heuvel, H. van den, Strik, H. (2006). Automatic Phonetic Transcription of Large Speech Corpora: a Comparative Study. In: *Proceedings of Interspeech - ICSLP*, Pittsburgh PA, USA, pp. 1085-1088.

Van Bael, C., Halteren, H. van (2006). On the Sufficiency of Automatic Phonetic Transcriptions for Pronunciation Variation Research. In: *Proceedings of Interspeech - ICSLP*, Pittsburgh PA, USA, pp. 717-720.

Van Bael, C., Boves, L., Heuvel, H. van den, Strik, H. (2006). Automatic Phonetic Transcription of Large Speech Corpora. In: *Proceedings of the Workshop on Annotation Science: State of the Art in Enhancing Automatic Linguistic Annotation*, held at the *International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, pp. 4-11.

Elffers, B., Van Bael, C., Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*. Internal report, Department of Language and Speech, Radboud University Nijmegen, the Netherlands.

Binnenpoorte, D., Van Bael, C., Os, E. den, Boves, L. (2005). Gender in Everyday Speech and Language: A Corpus-Based Study. In: *Proceedings of Interspeech - Eurospeech*, Lisbon, Portugal, pp. 2213-2216.

Van Bael, C., Heuvel, H. van den, Strik, H. (2004). Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora. In: *Proceedings of Interspeech - ICSLP*, Jeju, Korea, pp. 2793-2796.

Van Bael, C., Strik, H., Heuvel, H. van den (2004). On the Usefulness of Large Spoken Language Corpora for Linguistic Research. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 2135-2138.

Van Bael, C., Binnenpoorte, D., Strik, H., Heuvel, H. van den (2003). Validation of Phonetic Transcriptions Based on Recognition Performance. In: *Proceedings of Interspeech - Eurospeech*, Geneva, Switzerland, pp. 1545-1548.

Van Bael, C., Strik, H., Heuvel, H. van den (2003). Application-Oriented Validation of Phonetic Transcriptions: Preliminary Results. In: *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 1161-1164.

Van Bael, C., King, S. (2003). The Keyword Lexicon - An Accent-Independent Lexicon for Automatic Speech Recognition. In: *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 1165-1168.

Curriculum Vitae

Christophe Van Bael was born in Mortsel, Belgium on 12 December 1978. He enjoyed his primary education at the local Sint-Jozef primary school, and his secondary education at the nearby Sint-Gabriëlcollege in Boechout. From October 1996 to June 2001, he was enrolled as a graduate student in Germanic Philology at the Department of Linguistics and Literature of the University of Antwerp in Belgium. In June 2001, he graduated cum laude as Master of Arts, with majors in Dutch and English linguistics and computational linguistics. In September 2001 he left for Scotland, where he would follow a one-year Master course in Speech and Language Processing at the Department of Theoretical and Applied Linguistics of the University of Edinburgh. In September 2002, he graduated as Master of Science in Speech and Language Processing, and as European Master in Language and Speech. In October 2002, he took up a PhD position at the Department of Language and Speech of the Radboud University Nijmegen in the Netherlands. In addition to his PhD research he participated in the 'Leerwerktraject Basiskwalificatie Onderwijs', a professional training for lecturers. As part of this training, Christophe supervised bachelor students and organised courses and tutorials in the Faculty's Linguistics program. In September 2006, upon completion of his training, he received the 'Basiskwalificatie Onderwijs', a nationally acknowledged university-level teaching degree. Christophe is currently employed as a designer of spoken dialogue systems at the department Customer Contact Solutions at LogicaCMG in Nieuwegein in the Netherlands.