Using Rule Induction Techniques to Model Pronunciation Variation in Dutch

Véronique Hoste, Walter Daelemans, Steven Gillis

CNTS - Language Technology Group, University of Antwerp, Antwerp, Belgium

Abstract

In this paper, we present an inductive approach to the automatic extraction of knowledge about inter-regional pronunciation variation. We compare two different rule induction techniques, both popular in language engineering applications, viz. the rule sequence learner Transformation-Based Error-Driven Learning (TBEDL) (Brill, 1995) and the decision tree learner C5.0 (Quinlan, 1993). We investigate whether both techniques detect the same regularities and evaluate the extracted rules in terms of accuracy and in terms of linguistic relevance. As a case study, we apply the approach to Dutch and Flemish (the variety of Dutch spoken in Flanders, a part of Belgium), based on Celex and Fonilex, pronunciation lexica for Dutch and Flemish, respectively. Our main goal is to show that this approach allows the automatic acquisition of compact, interpretable translation rules between pronunciation varieties, on the basis of phonemic representations of words in both varieties (as output of phoneme recognition or, as in our case, on the basis of existing lexica). We also show that the observed differences coincide with the tendencies studied and described in linguistic comparative research of inter-regional pronunciation variation in standard Dutch.

Key words: Rule induction, machine learning, pronunciation variation

1 Introduction

A central component of speech processing systems is a pronunciation lexicon defining the relationship between the spelling and pronunciation of words. In most speech recognition and text-to-speech systems, the pronunciation lexicon contains only one possible pronunciation for a given word, which would be sufficient if words were always pronounced in the same way. However, pronunciation variation is a major problem in automatic speech recognition (see Strik and Cucchiarini (1999) for an overview of the literature on pronunciation variation). In order to reduce the error rates in ASR, pronunciation variation can be modeled by adding pronunciation varieties to the pronunciation lexicon. Pronunciation variation can vary from one single speaker pronouncing the same word in different ways (intra-speaker variation), to different speakers each with different pronunciations (inter-speaker variation). Alternate pronunciations, however, can also be caused by non-native speakers of a target language (see for example Livescu (1999) and Tomokiyo and Waibel (2001)) or can be caused by accents specific to a particular region. These *inter-regional varieties* of a language may differ considerably in their pronunciation. Examples are the pronunciation differences between different varieties of English, such as British, American and Australian English, which in turn all contain internal inter-regional pronunciation differences. Once a speaker from a particular region (e.g. USA) is detected, speech input and output systems should be able to adapt their pronunciation lexicon to this variety. Humphries et al. (1996), for example, modelled the variations in the pronunciations of vowels in Lancashire and Yorkshire accented speech in the British English. It is also on these inter-regional varieties, and more particularly the varieties within the Dutch standard language, that we focus in this paper.

Current approaches to modeling pronunciation variation can be divided into two groups depending on the source from which the information on pronun-

cation variation is derived. In the *data-driven* methods, the information on pronunciation variation is mainly obtained from the acoustic signal ((Cremelie and Martens, 1999), (Riley et al., 1999)). In the *knowledge-based* approach, on the other hand, the variation information is extracted from linguistic sources ((Kessens et al., 1999), (Adda-Decker and Lamel, 1999)) and the phonological tendencies described in the literature are used to derive different pronunciations of a given word form. A comparison between a knowledge-based approach and a data-oriented approach to modelling pronunciation variation in Dutch is given in Wester and Fosler-Lussier (2000). A similar comparison on the task of segmentation in German is decribed by Kipp et al. (1997). Both studies show that a data-driven approach can outperform a knowledge-based approach.

In the data-driven approach, DP-alignments are made between the transcripts of the acoustic signal and the transcriptions of single words in the lexicon. These alignments can then be used to derive formalizations. The information about pronunciation variation can be represented in terms of rewrite rules ((Cremelie and Martens, 1999)), decision trees ((Humphries et al., 1996), (Riley et al., 1999)) or neural networks ((Fukada et al., 1999)), etc.

In this paper, we aim at learning pronunciation rules from the data by using two rule induction techniques which have already been successfully applied in several language engineering applications. Our objective is similar to that of Cremelie and Martens (1999): trying to combine the advantages of both data-driven and knowledge-based systems. We apply this approach to Dutch: we try to extract the pronunciation differences between the Dutch as spoken in the Netherlands (Dutch) and the Dutch as spoken in Flanders (Flemish). We focus on the question whether it is possible to build in an inductive manner a compact set of pronunciation rules which reflects most of the differences between both varieties of Dutch. For the experiments, we apply two different rule induction methods, Transformation-Based Error Driven Learning (TBEDL) (Brill, 1995) and C5.0 (Quinlan, 1993) on two data sets, representing two different

varieties of Dutch, in order to extract information about the pronunciation differences between both variants. Whereas C5.0-like systems (decision trees) are often used in speech technology, TBEDL, developed in the context of language technology, has not been used before for this problem and seems to be well-suited with its ability for flexible generation of transformation rules. The key difference between both rule learners is that decision trees are applied to a set of non-interacting problem vectors. These problems are all solved independently. In Transformation-Based Error Driven Learning, on the other hand, rule sequences are learned. Rule sequence learning is applied to a sequence of interrelated problem instances, which are solved in parallel, by applying the rules to the entire corpus. This kind of learning allows the system to base its future learning on the application of earlier rules. The design of both systems is explained in more detail in Section 3. In Section 6, we investigate whether the rules produced by both rule learners also capture the same pronunciation differences. For a detailed discussion of the differences between both learning techniques, we refer to Ramshaw and Marcus (1994).

In our rule learning approach, it is feasible to generate more compact rule sets as it is possible to induce rules applying to sets of words rather than to individual words. The method allows the inductive construction of a rule set that can be integrated in a pronunciation lexicon of a speech processing system. Although we work here with existing, largely manually constructed, lists of pronunciations of words in both varieties, the approach should also be applicable to phonemic ASR output as well.

The design of this kind of regionally adaptable speech system assumes that the pronunciation differences are mostly systematic and can be modeled using rules. Besides focusing on the language engineering side of the study, we also discuss the quality and linguistic relevance of the extracted rules. We will show that the rules extracted by the rule-induction algorithms provide linguistic insights into the inter-regional pronunciation differences in the Dutch speaking

area and can also be useful in linguistic research.

In Section 2, we introduce the two pronunciation databases for Dutch that we used in the experiments. Section 3 gives a thorough description of the two rule induction techniques used in the experiments, viz. Transformation-Based Error-Driven Learning (TBEDL) and C5.0. In Section 4, the experimental setup is explained. This section is followed by a description of the results from both rule induction techniques. The rules learned by both methods are further analysed in Section 6. This section discusses the most important differences between Flemish and Dutch, starting from the first ten rules learned by both TBEDL and C5.0. In Section 7, we investigate the remaining pronunciation differences between both varieties which are not captured in the learned rule sets. In a final section, some concluding remarks are given.

2 Two Pronunciation Databases of Standard Dutch

For the study of the pronunciation of standard Dutch in Flanders and in the Netherlands, we made use of phonemic transcriptions of recordings of spoken standard Dutch. The success of such a data-oriented approach highly depends on the quality of the data being used in the experiments. When studying the pronunciation differences between two varieties of a language, this study should be based on a corpus which reflects the pronunciation differences between the speakers of both varieties. The corpus should reflect speech of a variety of speakers in a variety of speech situations, from more formal situations (e.g. speeches, texts read out loud) to spontaneous speech. For the study of inter-regional pronunciation in standard Dutch, there is up to now no corpus available that meets these standards. Therefore, we used for our rule induction experiments, both for Dutch and Flemish, two lexical databases representing these varieties. We based our study on the phonemic knowledge present in two databases, namely Celex and Fonilex, both representing one va-

riety of standard Dutch. Other researchers (Blancquaert, 1936) studying the differences in pronunciation of the standard Dutch between Flanders and the Netherlands have based their conclusions on the comparison of their own pronunciation with a source representing the other variant. Others make use of sound archives from the national radio or television stations with speech from politicians and policy makers (Cassier and Van de Craen, 1986), reporters (Van de Velde, 1996), and teachers of Dutch (van Hout et al., 1999).

Two lexica were used for this study, representing the Dutch and Flemish varieties. For Dutch Celex (release 2) (Baayen et al., 1993) was used and for Flemish Fonilex (version 1.0b) [1]. Since the phonemic transcriptions in both databases are based on different alphabets, we used IPA (International Phonetic Alphabet) to represent all phonemic transcriptions in order to avoid confusion.

The **Celex** database contains the phonemic transcription of more than 384,000 word forms as spoken in the Netherlands. The transcription of the stems of all words was done automatically and these transcriptions were then manually corrected. Inflections, derivations and compound forms were also automatically derived taking into account certain assimilation tendencies and these transcriptions were also checked by hand. For all word forms, only one possible transcription is given. In Celex, DISC [2] is used as the phonetic alphabet. Although a first version of the phonemic transcriptions was automatically obtained through the application of phonological rules, a large part of these transcriptions was corrected by hand.

The **Fonilex** pronunciation database contains the phonemic transcription of

[1] The Fonilex homepage is http://bach.arts.kuleuven.ac.be/fonilex/

[2] DISC is a computer phonetic alphabet is made up of distinct single characters, which a assigns one ASCII code to each distinct phonological segment. Further information on DISC can be found in Baayen et al. (1993).

the most frequent word forms of Dutch as spoken in Flanders. These word forms and their frequency information [3] are all taken from Celex. The word forms in Celex with a frequency of 1 and higher are included in Fonilex. From the list of words with frequency 0, only the monomorphematic words were selected. Celex only served as a basis for the list of word forms, not for the phonemic transcriptions. A first tentative representation of the pronunciation was provided by a memory-based learning automatic system for grapheme-to-phoneme conversion (Daelemans and van den Bosch, 1996) on the basis of a first batch of words manually transformed from Dutch to Flemish pronunciation. No phonological rules were used for this automatic process. This transcription was then verified manually and hand-corrected. Fonilex uses YAPA [4] instead of DISC as phonetic alphabet. The database consists of is a list of 218,113 entries, covering 205,216 different word forms with their Flemish pronunciation. The difference between these two figures is due to the double transcriptions for some word forms. The word "caravan", for instance, can be phonemically represented as /kɑrɑvɑn/ and as /kɛrɛvɛn/.

For both lexical databases, a first rough transcription was made automatically and this transcription was thoroughly corrected by hand. During this manual correction, the transcribers might have taken into account certain phonological rules they were familiar with from the phonological literature. No explicit rule sets, however, were used in this manual correction process. If these rules had existed, it would have been possible to compare these rules used for the creation of both databases with the rules detected by both rule induction

[3] The frequency information in Celex is based on the INL corpus, developed at the Institute for Dutch Lexicology (http://www.inl.nl). At the time information was extracted from it for Celex, the corpus contained 42,380,000 words, all taken from written resources of every kind.

[4] YAPA stands for "Yet Another Phonetic Alphabet". In appendix, a table is given of the different phonemes present in both databases and the mapping between the phonesets IPA, DISC and YAPA.

techniques (TBEDL and C5.0).

3 Rule Induction

Our starting point is the assumption that the differences in the phonemic transcriptions between Flemish and Dutch are highly systematic, and can be represented in a set of rules. These rules provide linguistic insight into the discrepancies between both varieties and can also be used to adapt pronunciation lexicons for Dutch automatically to Flemish and vice versa. In order to automatically detect these regularities, we decided to use two rule induction techniques, viz. the rule sequence learning technique Transformation-Based Error-Driven Learning (TBEDL) (Brill, 1995) and the decision tree learner C5.0 (Quinlan, 1993). These two rule induction algorithms were selected, since they have a completely different rule learning process. We were interested whether this would also lead to a different set of pronunciation rules, and if so, whether these two diverse perspectives could provide a useful complementary perspective on the task we are trying to solve.

In both Transformation-Based Error-Driven Learning (TBEDL) and C5.0, a set of rules is extracted from examples. The specification of a candidate set of rules on the basis of the examples is heuristic, and guided by compactness and accuracy criteria. In *classification rule induction*, represented by C5.0, each example consists of a pattern of feature values and a corresponding class symbol. The left-hand side of an induced rule is a logical combination of feature-value pairs (or sets thereof), and the right-hand side is a corresponding class. For each class, a set of rules is extracted. The *transformation-based learning* approach of TBEDL has a completely different flavor. While the left-hand side of an extracted rule still refers to feature values describing the triggering condition of the rule, the right-hand side describes a transformation of one class into another. Also, the examples are presented as two parallel

strings of feature values where one string has to be transformed into the other. In the remainder of this section, we will further discuss both rule induction algorithms. At the end of the section, we will investigate the main architectural differences between both rule learning approaches.

3.1 TBEDL

The first rule induction method, Transformation-Based Error-Driven Learning (Brill, 1995), has already been successfully applied to a number of natural language problems, such as part-of-speech tagging (Brill, 1992), prepositional phrase attachment (Brill, 1993) and syntactic parsing (Brill, 1996). All these tasks can be considered as **a sequence of interrelated problems**. In the case of part-of-speech tagging, for example, the part-of-speech of a given word largely depends on the parts-of-speech of the neighbouring words. Grapheme-to-phoneme conversion can also be considered as the resolution of a sequence of interrelated problems, since the phonemic representation of a given grapheme largely depends on the phonemic values of the surrounding phonemes. The TBEDL rule sequence learner is well adapted to a corpus that is inherently a sequence of interrelated problems. In TBEDL, all rules are applied to the entire corpus and this allows the learner to base its future learning for a given grapheme on the earlier application of rules for the neighbouring graphemes.

In Transformation-Based Error-Driven Learning (Brill, 1995), transformation rules are learned by comparing a corpus that is annotated by a so-called "initial state annotator", which provides a first rough transcription, to a gold standard correctly annotated corpus, which is called the "truth". In this learning process, the following procedure is followed:

- In each iteration, it is investigated for each possible rule how many mistakes of the "initial state annotator" can be corrected through application of that

rule.

- The rule which causes the greatest error reduction is retained.
- The rule causing the greatest error reduction is then the first rule of the rule list. This learned rule is applied to the output of the "initial state annotator" in order to bring that output closer to the "truth".

During each iteration, one rule is learned and directly applied to the output of the "initial state annotator". This learning process leads to an ordered list of transformation rules and learning stops when no errors can be corrected anymore.

For our experiments, we adapted the TBEDL learner to our needs. In this case, the task is to learn how to transform Celex representations into Fonilex representations (i.e., translate Dutch pronunciation into Flemish pronunciation) and vice versa. Both corpora serve as input for the "transformation rule learner" (Brill, 1995). In the two TBEDL experiments that were performed, both varieties function once as "truth" and once as text annotated by a so-called "initial state annotator". The learning process is repeated as long as rules can be found which detect pronunciation differences between both varieties of Dutch. The whole learning process leads to an ordered list of pronunciation rules. For both conversion processes, a different list of rules is learned. This approach to rule learning is similar to the one described by Cremelie and Martens (1999), who used an ordered set of positive and negative context dependent rewrite rules. Both approaches, however, use a different rule ordering strategy. In TBEDL, the rule ordering is "error-driven", which means that the rules are ordered according to the number of errors they correct. Applied to our experiments, this implies that the rules capturing the most pronunciation differences are highest in the rule list, followed by the rules capturing the second most differences, etc. The ordering of the rewrite rules in Cremelie and Martens (1999) is accomplished by arranging the rules according to their condition length: the most specific rules (the rules with the longest condition

Table 1

Graphemic (left-hand side of the slash) and phonemic (right-hand side of the slash) representation of the word "administreert" in Celex and Fonilex

Celex file	a/ɑ d/d m/m i/iː n/n i/iː s/s t/t r/r e/eː e/- r/r t/t
Fonilex file	a/ɑ d/d m/m i/ɪ n/n i/ɪ s/s t/t r/r e/eː e/- r/r t/t

part) are placed on top, whereas the most general rules are put at the bottom of the rule list.

Figure 1 shows the TBEDL learning process applied to the comparison of the Celex representation and the Fonilex representation. TBEDL takes as input two files, representing both varieties of Dutch. Each file contains the graphemic and phonemic representation of a Dutch word (as shown in Table 1 for the word "administreert" (Eng.: "administers")).

This learning process results in an ordered list of transformation rules which reflects the systematic differences between both representations. A rule consists of two parts: a *transformation* and a *triggering environment*. It can be read as: "change x into y (the transformation) in the following triggering environment". The following is a rule learned during our experiments:

/iː/ /ɪ/ NEXT1OR2OR3PHON /eː/

This rule can be divided into the transformation "change a Dutch tense /iː/ into a Flemish lax /ɪ/" and the triggering environment "if one of the three following Celex phonemes is a tense /eː/".

As already mentioned, it is investigated per iteration and for each possible rule/transformation how many pronunciation differences between both varieties of Dutch can be detected through application of that rule. The rule which covers most of the pronunciation differences is retained. In TBEDL, the rules are generated from a set of *transformation templates*. These transformation

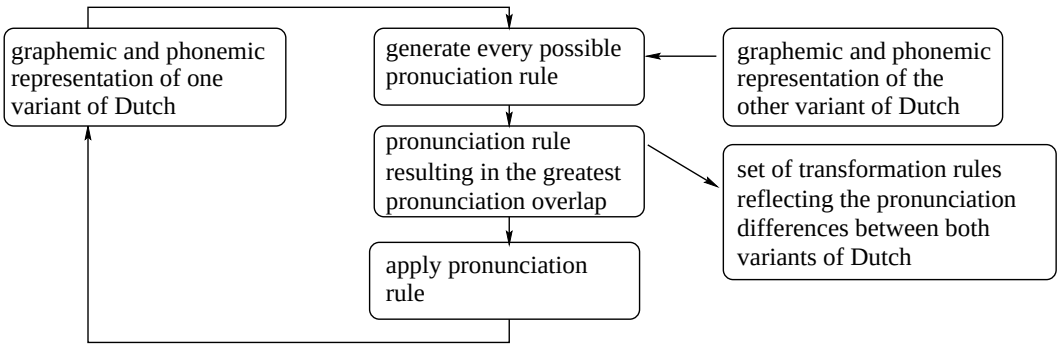

Figure 1. Architecture of the TBEDL learning process when learning Flemish and Dutch pronunciation differences

templates specify a small number of features or feature sets that are relevant for finding an appropriate rule. Rule sequence learning prespecifies in the templates essentially the full space of possible rules. This does not affect the power of the technique as long as the template set can be made rich enough to cover the patterns likely to be found in the data. In our task of deriving one variety of Dutch from the other variant, we decided to use the standard set of transformation templates provided in the Brill-learner (designed for part-of-speech tagging), which contains 26 different templates, as shown in Table 2. This set of transformation templates contains the graphemes and phonemes within a range of three positions to the left and three positions to the right of the target phoneme, e.g. "PREVPHON" (previous phoneme), "NEXT1OR2GRAPH" (one of the next two graphemes), "CURGRAPH" (current grapheme), "LBIGRAM" (left bigram), etc. The rules also take into account word boundary information. It is however also possible to define another set of templates (see for example Ramshaw and Marcus (1995)) and to extend the existing set with other mixes of grapheme and phoneme tests.

Per iteration, a rule is learned and this rule is applied to the phonemic transcriptions of one variety of Dutch in order to bring these phonemic transcriptions closer to those of the other variety of Dutch. The conversion of the Dutch pronuncation of the diminutive ending "tje" into its Flemish pronunciation is a nice illustration of this *rule sequence* learning approach. In TBEDL, two

Table 2

Set of transformation templates used in the learning process

Graphemes
CUR GRAPH
GRAPH AND 2 (AFT/BFR)
(NEXT/PREV) 1 GRAPH
(NEXT/PREV) 1 OR 2 GRAPH
(NEXT/PREV) 2 GRAPH
(L/R) BIGRAM
Phonemes
SURROUND PHON
(NEXT/PREV) 1 PHON
(NEXT/PREV) 1 OR 2 PHON
(NEXT/PREV) 2 PHON
(NEXT/PREV) 1 OR 2 OR 3 PHON
(NEXT/PREV) BIGRAM
Combining
GRAPH AND 2 PHON (AFT/BFR)
GRAPH (NEXT/PREV) PHON

rules are learned. First, the rule "change a /j/ into a /ʃ/ if the phoneme to the left is a /t/ and the phoneme to the right a /ə/" is applied to the whole corpus. In a second transformation rule learning round, the newly learned rule will base its learning on the application of the first rule. This leads to the rule

"omit the /t/ if the next phoneme is a /ʃ/". This *rule ordering* procedure implies also that it has to be taken into account that previously learned rules can be undone by a later rule (see also Roche and Schabes (1995)), as in the word "feuilleteer" (Eng.: "leaf through"). Celex provides the transcription /fœyjəteːr/ while Fonilex transcribes it as /føːjəteːr/. During learning, the transformation rule "change /œy/ into /øː/ if the preceding grapheme is an <e>" is learned. This results in the correct Fonilex-/føːjəteːr/. This transformation, however, is canceled by a later rule, which "changes /øː/ back into /œy/ if the following grapheme is an <i>." This leads again to the original Celex -transcription. C5.0 does not suffer from similar consequences of rule ordering and will correctly classify "feuilleteer". We will now proceed with the description of the C5.0 rule learning program.

3.2 C5.0

C5.0 (Quinlan, 1993) is the second rule induction program we used in the experiments. It has a rule learning approach, which is completely different from the TBEDL rule sequence learner, in which the corpus is considered as a set of interrelated problems. In C5.0, the decision trees are applied to **a set of non-interacting problem instances**. The trees are based on the unchanging features of the neighbouring graphemes and the decision tree learner is then forced to resolve the ambiguity at the neighbouring location as part of the rule for the primary site and it can only use as evidence the cases where the two occur together.

The program takes as input a set of examples consisting of a pattern of feature values and a corresponding class symbol. In our experiment of grapheme-to-phoneme conversion, the input pattern consists of graphemic and phonemic information. The task is defined as the conversion of fixed-size instances representing the focus grapheme ('fg') and focus phoneme ('fp'), with a certain

context to a class representing the target phoneme, as shown in Table 3. To generate the instances, windowing is used according to a technique proposed by Sejnowski and Rosenberg (1987). This means that we move a window over the words, where the window spans several letters. In the example presented in Table 3 we made use of a context of three phonemes preceding (indicated by fp-1, fp-2, and fp-3) and three phonemes following (fp+1, fp+2, fp+3) the focus phoneme. The graphemes are indicated by an 'fg' followed by a number indicating the position of the grapheme. "=" is used as boundary symbol. The '-' is an alignment symbol (see Section 4 for further information on alignment), which is used whenever the phonemic form is shorter than the graphemic representation.

C5.0 generates a classifier in the form of a decision tree. In the building of a decision tree, an elementary predicate is selected at each step to split a single leaf node, meaning that it is applied only to those training instances associated with that particular branch of the tree. The decision tree classifies a case starting at the root of the tree and then moves through the tree until a leaf node (associated with a class) is encountered. This decision tree can also be converted into a set of production rules. Making rules for each leaf of the tree would not generate a rule set that is easier to understand. Therefore, the rules are first generalized by deleting irrelevant conditions. In a second step, all simplified rules for a given class are checked and the rules which do not lead to a greater accuracy are removed. This leads to a more compact and understandable rule set.

The rules have the form "L -> R", in which the left-hand side is a logical combination of feature-value pairs (or sets thereof), and the right-hand side a corresponding class. When classifying a new unseen case, the list of rules is examined to find the first rule whose left-hand side satisfies the case. In order to produce more concise decision trees and rules, a value grouping method is invoked, which collapses different values for a (graphemic or phonemic)

Table 3

The instances generated for the word "administreert" (Eng. "administers") for a C5.0 experiment converting Celex pronunciation into Fonilex pronunciation.

graphemic representation			phonemic representation			class
left	**fg**	right	left	**fp**	right	
===	a	dmi	===	ɑ	dmiː	ɑ
==a	d	min	==	d	miːn	d
=ad	m	ini	=ɑd	m	iːniː	m
adm	i	nis	ɑdm	iː	niːs	ɪ
dmi	n	ist	dmiː	n	iːst	n
min	i	str	miːn	iː	str	ɪ
ini	s	tre	iːniː	s	treː	s
nis	t	ree	niːs	t	reː-	t
ist	r	eer	iːst	r	eː-r	r
str	e	ert	str	eː	-rt	eː
tre	e	rt=	treː	-	rt=	-
ree	r	t==	reː-	r	t==	r
eer	t	===	eː-r	t	===	t

feature into subsets. This leads to subtrees or rules associated with a subset of values rather than with a single value. These attribute value groups have the form "A in $\{V_1, V_2, ...\}$", e.g. fp-1 in {aː, eː, iː, oː, yː}. The method Quinlan (1993) uses to find groups of attribute values, is based on iterative merging of value groups. The initial value groups are then the individual values of a

given attribute. In each cycle, the consequences of merging every pair of value groups are evaluated.

The following rule is an example of the rules learned by C5.0 during the experiments for the conversion of Dutch into Flemish pronunciation.

(7688/30, lift 112.1)
fp-1 in {=, oː, juː}
fp in {x, ɡ}
-> class ɣ [0.996]

The rule shows the conversion of the Dutch phonemes /x/ and /ɡ/ into the Flemish phoneme /ɣ/ in case of the directly preceding phonemes = (word boundary), /oː/ and /juː/. At the top of the rule the number of training cases covered by the rule (7688) is given together with the number of covered cases that do not belong to the class predicted by the rule (30). The "lift" (112.1) is the estimated accuracy of the rule divided by the prior probability of the predicted class. The accuracy of this rule on the train set is given at the right-hand side of the rule (99.6%).

TBEDL and C5.0 generate two different types of rules. The transformation rules, which are learned in TBEDL, focus on the differences between Dutch (Celex) and Flemish (Fonilex) and only the differences between those both varieties are captured in the rules. The C5.0 production rules, on the other hand, also describe the overlapping phonemes between Celex and Fonilex. In order to make the type of task of C5.0 comparable with the transformation based approach used by TBEDL, we changed the output class to be predicted by C5.0 to '0' when the Celex and Fonilex phoneme were identical (i.e. no change) and to the target phoneme when Celex and Fonilex differed. A second difference between the output of both rule learners is that TBEDL appears to have much less power to create complex combined rules than do decision

Table 4

The use of compounds in “taxi”.

Word form	t	a	x	i
Without compounds	t	ɑ	ks	iː
With compounds	t	ɑ	X	iː

trees. Because rule sequence learners are more limited in terms of the connections between rules that they can construct during learning, they must begin mith more complex predicates built into their rule templates. Decision trees, on the other hand, synthesize complex rules from elementary predicates by inheritance.

In Section 6, we investigate whether the rules produced by both rule learners also capture the same pronunciation differences.

4 Experimental Setup

4.1 Preprocessing

Before presenting the data to TBEDL and C5.0, two preprocessing steps were taken, viz. the insertion of compound symbols and alignment. Compound phonemes are used whenever graphemes map with more than one phoneme, as in the word “taxi”, in which the <x> is phonemically represented as /ks/ in /tɑksiː/. Other examples are gen**i**aal (<i> transcribed as /ɪj/), aster**oi**de (<oi> transcribed as /ɔwi/) and empl**oy**e (<oy> transcribed as /wɑj/). This problem is solved by defining a new phonemic symbol that corresponds to the two phonemes, as indicated in Table 4. Roughly 30 phonemic combinations in both Celex and Fonilex were replaced by compound symbols.

Furthermore, alignment is required since the phonemic representation and

Table 5

Alignment of the word “aalmoezenier” (Eng.: “chaplain”).

a	a	l	m	o	e	z	e	n	i	e	r
aː	-	l	m	uː	-	z	ə	n	iː	-	r

the spelling of a word often differ in length (Daelemans and van den Bosch, 1996). Therefore, the phonemic symbols are aligned with the graphemes of the written word form. In case the phonemic transcription is shorter than the spelling, null phonemes ('-') are used to fill the gaps, as shown in Table 5. In this experiment, alignment was performed for the graphemic and phonemic representations of Celex and for those of Fonilex.

4.2 Train and Test Material

For both sets of experiments, viz. the conversion of Dutch into Flemish and vice versa, the orthographic word form and its phonemic transcription were used. The corpus used for the experiments contains the word forms present in both databases. Only one transcription per word form is taken into account and for the word forms which have more than one possible transcription, only one transcription is randomly selected. Words of which the phonemic transcription is longer than the orthography and for which no compound phonemes[5] are provided are omitted, e.g. “b'tje” (phonemically: /beːtjə/). After omission of these double transcriptions, the corpus consists of 202,136 different word forms or 1,972,577 phonemes. This is the corpus we will use for the experiments.

[5] In order to avoid a large set of phonemes, we only created compound phonemes if the phonemic combinations occured more than 10 times in the corpus. It would however also have been possible to create compounds every time a grapheme maps with more than one phoneme.

For both varieties of Dutch, both algorithms are trained on a train set and evaluated on a test set. 10% (197,257 phonemes) of the data set is used for evaluation.

For the **TBEDL** experiment, the remaining 90% is used for learning the transformation rules. For this experiment, a threshold of 15 errors is specified, which means that learning stops if the error reduction lies under that threshold. Due to the large amount of training data, this threshold is chosen to reduce training time. It is however also possible to lower this threshold. As described in Section 3, this approach is central to TBEDL: during each iteration, all possible rules are applied and the rule causing the greatest error reduction is retained. The term 'error', however, should not be taken literally in our experiments. Setting the threshold in our experiments to 15 errors, for instance, means that learning stops if less than 15 instances out of the 1,775,320 training instances are covered by a certain Celex-to-Fonilex or Fonilex-to-Celex transformation rule. Even a substantial increase of this threshold would still result in a set of rules describing the most frequent differences between both pronunciation variants.

The **C5.0** learning algorithm was also trained on the same 90% train set of 1,775,320 instances, but this experiment was computationally too expensive. Therefore, a new training experiment was conducted with 50% (887.660 cases) of the original training set as new training set. A decision tree model was built from the training cases. The tree was then converted into a set of production rules and these rules were evaluated on the 10% test set.

5 Experimental Results

5.1 Overlap between both Pronunciation Varieties

In order to detect the distribution of the pronunciation differences between both varieties of standard Dutch, we calculated the overlap and the differences between both varieties on the word and on the phoneme level. These calculations were performed on the 10% test set, in order to allow for comparison with the percentages obtained after rule learning. When comparing our Celex and Fonilex corpus, an initial overlap of 59.1% on the word level and 92.8% on the phoneme level could be observed. Consonants (96.0%) and diphthongs (99.8%) are highly overlapping (see Table 6). The overlap for the vowels is lower: 85.6%.

Table 6
Initial overlap between Celex and Fonilex

Words	Phonemes	Consonants	Vowels	Diphthongs
59.07%	92.77%	95.95%	85.58%	99.76%

In the following experiments, we tested whether rule induction techniques were able to learn to adapt Dutch (Celex) pronunciations to Flemish (Fonilex) when trained on a number of examples and vice versa. A successful rule learning procedure would then be able to learn the pronunciation differences between both varieties of Dutch. By using Transformation-Based Error-Driven Learning and C5.0, we looked for the systematic differences between Dutch and Flemish. If this method of "translating" one pronunciation variety into the other would be successful, this means that it is possible to capture the pronunciation differences between both varieties of standard Dutch in a *compact set of rules*.

Two sets of experiments were performed using TBEDL: one for the conversion of Celex into Fonilex and a second one to convert Flemish (Fonilex) into Dutch (Celex) pronunciation. This resulted in about 450 transformation rules for the conversion of Celex into Fonilex and into about 250 rules for the conversion in the opposite direction. This large difference in the number of rules can be explained by the fact that the Flemish corpus contains more pronunciation variation, such as the use of nasal sounds in loan words, than the Dutch corpus. In the word "grandeur" (Eng.: "splendor"), for instance, the <n> is represented as the phoneme /˜/ (/grɑ˜dəːr/) in Fonilex, which is only used to describe the pronunciation of the <n> in loan words. This Fonilex /˜/ is learned in different transformation rules taking into account both graphemic and phonemic information. In Celex, the loan words are described by the phonemes provided for the regular Dutch words, in this case /n/ (/grɑndøːr/). This Dutch /n/ can be derived from the Flemish /˜/ through one simple rule: "change a Flemish /˜/ into a Dutch /n/ if the currrent grapheme is a <n>".

In order to detect the most frequent pronunciation differences between both varieties of Dutch, we plotted the number of transformation rules against the accuracy of the conversion between Celex and Fonilex (Figure 2). A 100% accuracy would mean that all pronunciation differences between Dutch and Flemish are captured in the transformation rules. Both plots clearly show the same tendencies in the accuracy percentages both on the word and the phoneme level. For both deriving Celex transcriptions from Fonilex transcriptions and vice versa, we can see that especially the first 50 rules lead to a considerable increase of performance. For the conversion of Celex transcriptions into Fonilex transcriptions, performance increases from 59.1% to 79.4% on the word level and from 92.8% to 97.0% on the phoneme level when ap-

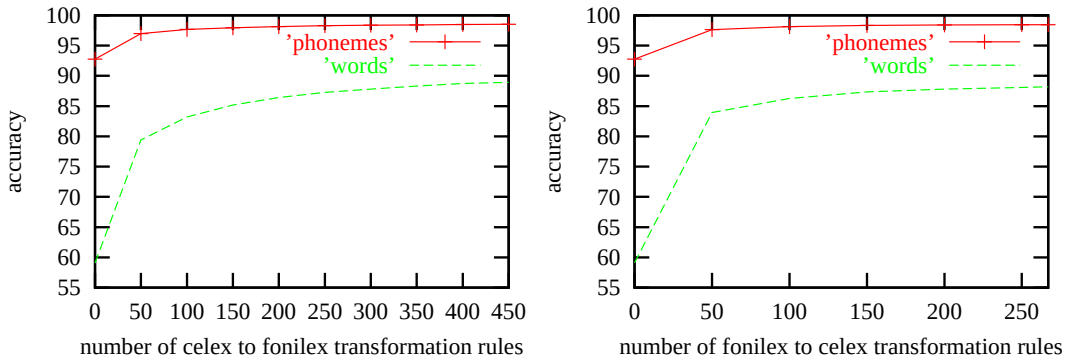

Figure 2. Description of the accuracy of the word and phoneme level in relation to the number of TBEDL transformation rules

plying the first 50 rules, which indicates the high applicability of these rules. For the Fonilex to Celex conversion process, the increase is even larger: the initial accuracy increased to 83.0% on the word level when applying those first 50 rules. For the phonemes, the accuracy increased to 97.7%. Afterwards, the increase of accuracy is more gradual: from 79.4% to 89.0% (words) and from 97.0% to 98.5% (phonemes) for the derivation of the Flemish pronunciation. And for the derivation of the Dutch pronunciation, accuracy increases from 83.0% to 88.2% (words) and from 97.6% to 98.5% (phonemes).

These scores indicate that we were able to successfully convert one variety into the other. Moreover, even a restricted rule set of 50 transformation rules could already detect the most frequent pronunciation differences between Dutch and Flemish. So, the application of TBEDL on both data sets has led to a set of valuable transformation rules capturing most of the pronunciation differences between both varieties of Dutch.

5.3 Classification-Based Rule Learning

In the C5.0 experiment, a decision tree model was built from the training cases and this tree was then converted into a set of production rules. The tree was converted into a set of 709 rules for the conversion of Celex transcriptions

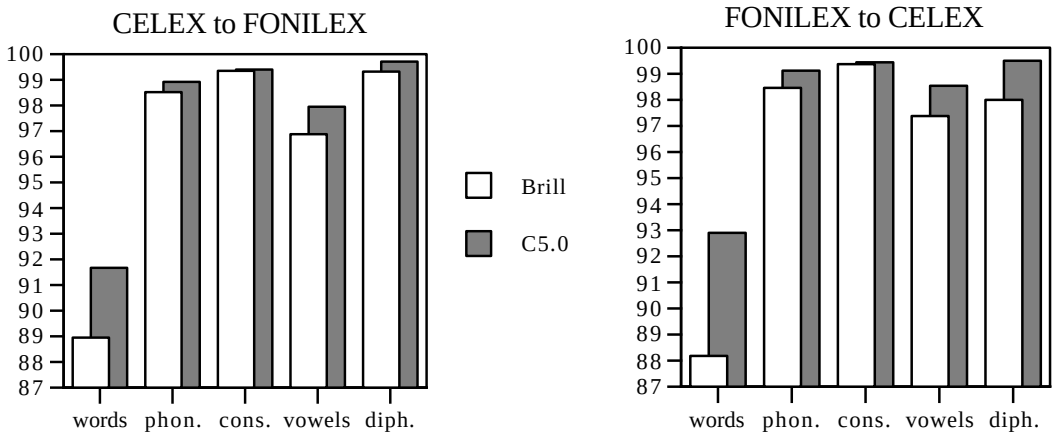

Figure 3. Accuracy for the words, phonemes, consonants, vowels and diphthongs after application of all TBEDL transformation rules and C5.0 production rules into Fonilex transcriptions. When learning Celex pronunciation, 658 rules were learned. These production rules were applied to the original 10% test set we also used in the Brill experiment. The output class to be predicted by C5.0 was either '0' when the Celex and Fonilex phoneme were identical (i.e. no change), or the target phoneme when Celex and Fonilex differed. Learning Dutch pronunciation resulted in 193 0-rules. For Flemish, 207 0-rules were learned. The fact that C5.0 generates more rules than TBEDL, could be explained by the nature of both algorithms. In TBEDL, the rule ordering implies that intermediate results in classifying one object can be used for the classification of other objects, which is not the case in a classification-based approach, such as C5.0.

Figure 3 gives an overview of the accuracy on the word and phoneme level for both conversion processes after application of the rule induction techniques. A comparison of these results shows that, when evaluating both TBEDL and C5.0 on the same test set, the transformation rules learned by the Brill-tagger have a higher error rate, even though C5.0 is only trained on half the data used by TBEDL. In order to determine whether or not these differences in accuracy between both algorithms were statistically significant, we performed a z-test (a variant of the t-test)(see Freedman et al. (1991) for a description

of this test). According to this test, all differences, both on the word and the phoneme level, are statistically significant.

When learning the Flemish pronunciation, an accuracy of 89.0% on the word level is reached when applying all transformation rules. The application of the C5.0 production rules leads to a 91.7% word accuracy ($z=12.27$, $p<0.001$). On the phoneme level, the use of the Brill-tagger leads to a 98.5% accuracy and the use of the C5.0 production rules leads to a 98.9% accuracy ($z=14.62$, $p<0.001$). The same tendency can be observed, when learning the Dutch pronunciation. After application of the transformation rules, there is an 88.2% accuracy on the word level. When applying all C5.0 rules, there is a 92.9% accuracy on the word level ($z=20.71$, $p<0.001$). With regard to the phonemes, a 98.5% accuracy is observed when using TBEDL and a 99.1% when using C5.0 ($z=21.92$, $p<0.001$). In both learning experiments, C5.0 also has a slightly lower error rate for the consonants, vowels and diphthongs. These results indicate that the rules learned by C5.0 more accurately describe the pronunciation differences between the Flemish and Dutch variety of the Dutch standard language.

A comparison of the initial differences between both varieties of Dutch and the final accuracy after application of the rules shows how many differences on the word and phoneme level can be predicted by the Brill and the C5.0 rules. For the conversion of Celex into Fonilex, we see that it is possible to learn transformation rules which predict 73% of these differences at the word level and 79.5% of the differences at the phoneme level. The C5.0 rules are roughly 6% more accurate: 79.7% (words) and 85.1% (phonemes). For the conversion of Fonilex into Celex, the transformation rules predict 71.1% of the initial differences at the word level and 78.6% of the differences at the phoneme level. The C5.0 rules outperform the Brill-rules: 82.7% (words) and 87.8% (phonemes).

We can conclude that it is indeed possible to reliably 'translate' Dutch into

Flemish and vice versa by using these rule-induction techniques. A possible explanation for the fact that the rules learned by C5.0 more accurately describe the pronunciation differences between Flemish and Dutch, could be that the set of transformation templates in TBEDL is not refined enough to capture the important patterns in the data.

6 Rule Analysis

In this Section, we analyse in more detail the rules learned by both rule induction techniques. We compare the initial differences between both pronunciation varieties (reported in Table 6) with the results obtained after application of the rules. For this comparison, we focus on the rules learned during the conversion of Fonilex into Celex. We discuss the rules learned for the consonants, vowels and diphthongs separately and compare these results with the observations made in phonological literature on standard Dutch. Linguistic literature, such as Booij (1995), Van de Velde (1996) and De Schutter (1978) indicates tendencies such as voicing and devoicing on the consonant level and the confusion of tense and lax vowels as important differences between Dutch and Flemish. We will discuss these tendencies in more detail and investigate whether they can also be induced from the two pronunciation lexica.

For this analysis, we make use of the first ten transformation rules that are learned for the conversion of Fonilex (Flemish) pronunciation into Celex (Dutch) pronunciation (see Table 7). We will discuss these rules in more detail and compare them with the 10 non-0 production rules, which most reduce the error rate. This analysis will also reveal that the rules produced by both learning techniques mainly capture the same pronunciation differences.

Table 7

Overview of the first ten transformation rules for the conversion of Fonilex into Celex. The "STAART" in the first rule indicates a word boundary in TBEDL.

Fonilex	Celex	Triggering environment	Example
ɣ	x	PREVPHON STAART	Fon.: ɣəlɛikaːrdəx (Eng. equal) Cel: xəlɛikaːrdəx
ɪ	iː	NEXT 1 OR 2 GRAPH e	Fon.: mʉltɪplɪseːr (Eng. multiply) Cel.: mʉltiːpliːseːr
tʃ	j	NEXT 1 OR 2 OR 3 PHON ə	Fon.: aːjʃə (Eng. stroke) Cel.: aːjtjə
-	t	NEXT BIGRAM jə	
ɑ	aː	NEXT 1 OR 2 GRAPH i	Fon.: pɑpiːr (Eng. paper) Cel.: paːpiːr
ɔ	oː	NEXT 1 OR 2 GRAPH e	Fon.: kɔntrɔleːrbaːr (Eng. verifyable) Cel.: kɔntroːleːrbaːr
ɪ	iː	NEXT 2 GRAPH i	Fon.: rɪviːr (Eng. river) Cel.: riːviːr
ɔ	oː	NEXT 2 GRAPH i	Fon.: pɔliːtɪkɑ (Eng. politician) Cel.: poːliːtiːkaː
ɪj	iːj	CUR GRAPH i	Fon.: aːtrɪjʉm (Eng. atrium) Cel.: aːtriːjʉm
ɑ	aː	GRAPH AND 2 AFT a e	Fon.: ɑkɑdemiː (Eng. academy) Cel.: aːkaːdemiː

The figures in Table 6 show that Dutch and Flemish are highly similar on the consonant level (96.0%). The aim of both rule learning procedures was to detect the remaining differences between both varieties of Dutch and then to apply those rules on the 10% test set. This would then enable us to convert Celex pronunciation into Fonilex pronunciation and vice versa. For the conversion of Fonilex consonants into Celex consonants, the application of both rule sets leads to a 99.4% accuracy. This means that, through rule induction, we are able to capture nearly all pronunciation differences on the consonant level.

Let us now have a closer look at the different consonant rules that were learned by both techniques. An analysis of the differences on the consonant level shows that 60% of these differences concerns the alternation between voiced and unvoiced consonants. In the word "gelijkaardig" (Eng.: "equal"), for example, we find /xəlɛikaːrdəx/ with an initial voiceless velar fricative in Dutch and /ɣəlɛikaːrdəx/ with a voiced velar fricative in Flemish. The word "machiavellisme" (Eng.: "Machiavellism") is pronounced with a voiceless /s/ in Dutch (/mɑɣiːjaːvɛlɪsmə/) and with a voiced /z/ in /mɑkɪjɑvɛlɪzmə/ in Flemish. These differences between voiced and unvoiced consonants have also been described in the phonological literature: the different realizations of the dental fricatives /s/ and /z/, the labiodental fricatives /f/ and /v/ and the velar fricatives /x/ and /ɣ/ have also been studied and described by Cassier and Van de Craen (1986), Van de Velde (1996), Smakman and van Bezooijen (1999) and others.

A closer look at the confusion matrix in Table 8 shows that, among these alternations between the voiced and unvoiced consonants, especially the alternation between /x/ and /ɣ/ is very frequent. This alternation is also the subject of the first transformation rule that is learned in the conversion of

Table 8

Confusion matrix for the voiced and unvoiced consonants in the test corpus.

Celex	Fonilex							
	t	d	f	v	s	z	x	ɣ
t	14774	127						
d	30	6516						
f			2438	14				
v			24	3219				
s					10498	327		
z					57	1992		
x							2743	1880
ɣ							92	2373

Fonilex pronunciation into Celex pronunciation: "/ɣ/ changes into /x/ in case of a word boundary one position before". is learned. This alternation is also described in the top ten of the C5.0 production rules, which most reduce error rate. The C5.0 rule also shows that the rule is successfully applied to 7638 instances in the train set, yielding a 99.3% accuracy on that train set.

(7638/56, lift 113.3)

fg-1 in {=, E, V, R}

fp = ɣ

fp+1 in {=, aː, x, j, ə, t, d, n, tʃ, s, k, l, b, ɛ, ʒ, ɛi, r, (...)}

-> class x [0.993]

Table 7 reveals another frequent pronunciation difference on the consonant level, viz. the use of palatalization in Flemish. This is also extensively described in the phonological literature (e.g. van Hout et al. (1999)). For the diminutive word "aaitje" (Eng.: "stroke"), for instance, Fonilex uses the palatalized form /aːjtʃə/ instead of the Celex form /aːjtjə/. Two Brill rules make this change possible. When learning the Dutch pronunciation of the diminutive ending "tje", /tʃ/ first changes into /j/. And in a second step, a /t/ is added in front of the bigram /jə/. This change is also described in the top 10 of C5.0 rules.

6.2 Vowels

Before the rule learning, 85.6% of the vowels in both pronunication lexica were pronounced the same. Rules were then learned for the remaining differences between both varieties of Dutch and these rules were then applied to the 10% test set. For the conversion of Fonilex consonants into Celex consonants, the application of the TBEDL transformation rules leads to a 97.4% accuracy and the application of the C5.0 rules leads to a 98.5% accuracy. So, a large part of the initial pronunciation differences between Dutch and Flemish are captured by those rules.

An analysis of the differences at the vowel level between Celex and Fonilex shows that 96% of the differences between the Flemish and Dutch vowel pronunciations concerns the use of a lax vowel instead of a tense vowel for the /iː/, /eː/, /aː/, /oː/ and /uː/. This alternation is illustrated by the confusion matrix in Table 9, which clearly shows that tense Celex vowels not only correspond with tense, but also with lax vowels in Fonilex. These differences in vowel quantity (duration) between Dutch and Flemish are also described in linguistic literature (e.g. Goossens (1973)).

Table 9

Confusion matrix showing the use of Flemish lax and tense vowels given the Dutch tense vowels.

Celex	Fonilex									
	iː	yː	eː	aː	oː	ɪ	ʉ	ɛ	ɑ	ɔ
iː	2302					2632				
yː		387					519			
eː			4384					993		
aː				3507					1797	
oː					2546					1606

This transition from lax vowels (such as /ɪ/, /ɑ/, /ɔ/, /ɛ/) into the corresponding tense vowels (/iː/, /aː/, /oː/, /eː/) in a certain triggering environment is clearly shown in seven out of the ten rules in of Table 7 (see transformation rules 2, 5, 6, 7, 8, 9, 10). An example is the word "multipliceer" (Eng.: "multiply") which is transcribed as /mʉltiːpliːseːr/ in Celex and as /mʉltɪplɪseːr/ in Fonilex. A closer look at the ten most important C5.0 production rules shows the same tendency: seven out of ten rules describe this alternation between a tense and a lax vowel. E.g.

(1440/5, lift 408.1)

fg+1 in {g, j, t, n, d, s, k, l, b, r, m, z, p, c, v, f, x}

fg+2 in {e, i, u}

fp = ʉ

fp+2 in {j, ɛ, eː, ɑ, uː, iː, ʉ, ɔ, aː, ɪ, oː, yː, iːj, ɪj, ɛj, ɒː, eːj, ɔʊ, oːʊ, ɑj}

-> class yː[0.943]

With regard to the pronunciation of diphthongs in Flemish and Dutch, both lexical databases are highly similar. Table 6) shows that 99.8% of the diphthongs in the 10% test set is pronounced the same in both varieties of Dutch. After application of the rules, this score slightly decreased, which could be caused by the overgeneralization of some of the learned rules. When learning the Celex pronunciation of diphthongs, for example, a 98% is obtained after application of the TBEDL transformation rules and a 99.5% after application of the C5.0 production rules. In the top ten rule set of both rule induction algorithms, there is no rule describing the differences on the diphthong level, which is not surprising given the large initial similarity between both lexica. Also few linguistic studies mention differences on the diphthong level, e.g. Van de Velde (1996) observes regionally different realizations of the diphthong <ei> but also mentions that these deviating realizations are not accepted in the Dutch standard language.

All the top ten rules described in this section, capture the most frequent phonemic differences between Dutch and Flemish. They are only a fraction of the rules learned by both rule-induction methods. An example of a less frequent, but very consistent rule that is learned by both techniques is the t-deletion in Flemish in words ending on <tie>, e.g. "politie" which is pronounced as /poːliːtsiː/ in Dutch and as /pɔliːsiː/ in Flemish. This difference has also been observed and described by e.g. Van Haver (1972) in French loan words after a vowel and after the graphemes <n> or <r>.

The overlap between the rules found by both rule induction techniques based on two lexica of spoken Dutch and the tendencies described in phonological literature shows that rule induction is a valuable instrument to detect and model pronunciation variation in a language. Both on the consonant and vowel level, a lot of similar tendencies are observed when comparing the inductively ob-

tained rules on the basis of two lexical data bases (Celex and Fonilex) and the observations described in comparative studies on inter-regional pronunciation variation in the Dutch standard language. The rule induction methods have the additional advantage that an exhaustive list of pronunciation differences is induced, also capturing the less frequent differences, whereas linguistic studies mostly limit themselves to a predefined set of phonological variables. A disadvantage of the corpus-based techniques is that they depend on the phonetic encodings of the lexica being used, which do not always capture all possible realizations of a sound. Both lexica are only provided with a phonemic or broad phonetic transcription, which is not suitable to describe phonetic differences, which require a narrow phonetic transcription. A further advantage of both rule induction techniques is that they can also track down systematic false or deviating transcriptions in the corpora.

7 Remaining Pronunciation Differences

Besides the systematic phonemic differences between Flemish and Dutch, there are a number of unsystematic differences between both databases. In this section, we are concerned with the remaining differences after application of all rules. Therefore, a manual analysis of these remaining differences was done. As a test case, we studied the conversion of Dutch into Flemish. This showed that the explanation of these remaining errors is twofold:

- A first reason is that no rule is available for less frequent cases. The rules are induced on the basis of a sufficiently large frequency effect. This leads to no rule at all for less frequent phonemes and phoneme combinations and also for phonemes which are not always consistently transcribed. Examples of infrequent words are loan words, such as “points” and “panty’s” or the loan sound /˜/ which only appears in Fonilex.
- A second reason is that rules will over-generalize in certain cases. The con-

fusion matrix for vowels in Table 9 clearly indicates the tendency to use more lax vowels in Flemish. This leads to a number of Brill and C5.0 rules describing this tendency. But both rule learners commit errors. A closer investigation of the errors committed by the Brill-tagger, for example, shows that 41.7% of the errors concerns the use of a wrong vowel. In 25.0% of the errors committed on the phoneme level, there was an incorrect transition from a tense to a lax vowel, as in "antagonisme" (Eng.: "antagonism") where there was no transition from an /oː/ to an /ɔ/. In 16.8% of the errors, a tense vowel is erroneously used instead of a lax vowel, as in "affiche" (Eng.: "poster") where an /ɪ/ is used instead of a (correct) /i/. Difficulties in the alternation between voiced an unvoiced consonants account for 6.3% of the errors on the phoneme level. E.g. in "administratie" the /t/ was not converted into /d/.

In order to analyze why C5.0 performs better on our task than TBEDL, a closer comparison was made of the errors exclusively made by the Brill-tagger and those exclusively made by C5.0. However, no systematic differences in errors were found which could explain the higher accuracies when using C5.0.

8 Concluding Remarks

In this paper, we combined the advantages of a data-driven and knowledge-based approach to the modeling of pronunciation variation in Dutch. We investigated whether two rule induction algorithms, viz. the rule sequence learning algorithm Transformation-Based Error-Driven Learning (Brill, 1995) and the decision tree learner C5.0 (Quinlan, 1993), which are both very popular in language engineering applications, could also produce compact and interpretable translation rules between two pronunciation varieties (Flemish and Dutch) in the Dutch standard language. We showed that both methods allow the inductive construction of a rule set. We also showed that, in spite of the

architectural differences between both learning techniques, they mainly capture the same pronunciation differences. Although we work here with existing, largely manually constructed, lists of pronunciations of words in both varieties, the approach should be applicable to phonemic ASR output as well. We also showed that the observed differences coincide with the tendencies studied and described in linguistic comparative research of inter-regional pronunciation variation in standard Dutch.

A quantitative and qualitative analysis was given of the phonemic differences discovered by both rule induction techniques when trained on the Celex database (Dutch) and the Fonilex database (Flemish).

Studying the overall accuracy in predicting the pronunciation of a Flemish word pronunciation from the Dutch pronunciation, a ca. 89% accuracy for TBEDL and 92% for C5.0 (ca. 99% at phoneme level for both) is obtained. For the conversion of Flemish into Dutch pronunciation, the same tendencies can be observed: an overall accuracy of 88% is reached in predicting the pronunciation of a Dutch word when applying the transformation rules. When applying all C5.0 rules, 93% of the words are pronounced the same in Dutch and Flemish. With respect to the phonemes, a 98% accuracy is observed when using TBEDL and a 99% when using C5.0. The accuracies of both learning techniques indicate that it is indeed possible to reliably convert Dutch into Flemish and vice versa by making use of rule induction. Moreover, the use of these rule-induction techniques can be an appropriate method for adapting pronunciation databases of one variety automatically to the other variant. Several pronunciation varieties of a language could be stored as compact sets of transformation rules in a speech recognition or synthesis system. Ideally, the extraction of the rules could be done using phoneme recognition systems, or using annotated corpora (annotation is cheaper than manual rule development).

A qualitative analysis of the first ten rules produced by both methods, suggested that both TBEDL and C5.0 extract valuable rules describing the most important linguistic differences between Dutch and Flemish on the consonant and the vowel level. The pronunciation differences extracted both by TBEDL and C5.0 largely coincide with the tendencies studied and described in linguistic comparative research of inter-regional pronunciation variation in standard Dutch. Linguistic literature also indicates tendencies such as voicing and devoicing on the consonant level and the confusion of tense and lax vowels as important differences between Dutch and Flemish. The rule-induction methods described in this paper are not only a valuable instrument to extract the most salient pronunciation differences between both varieties of Dutch. They have the additional advantage that an exhaustive list of pronunciation differences is induced, also capturing the less frequent differences, whereas linguistic studies mostly limit themselves to a predefined set of phonological variables.

Acknowledgements

This research was partially funded by the FWO projects Linguaduct and Prosit and the IWT project CGN (Corpus Gesproken Nederlands).

References

Adda-Decker, M., Lamel, L., 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication 29 (2), 83–98.

Baayen, R., Piepenbrock, R., van Rijn, H., 1993. The CELEX lexical data base on CD-Rom. Philadelphia, PA: Linguistic Data Consortium.

Blancquaert, E., 1936. Noord- en zuidnederlandsche schakeeringen in de

beschaafd-nederlandsche uitspraak. In: Verslagen en Mededelingen van de Koninklijke Vlaamse Academie voor Taal en Letterkunde. pp. 597–612.

Booij, G., 1995. The phonology of Dutch. Oxford: Clarendon Press.

Brill, E., 1992. A simple rule-based part-of-speech tagger. In: Proceedings of the 3rd Conference on Applied Natural Language Processing. pp. 152–155.

Brill, E., 1993. Automatic grammar induction and parsing free text: A transformation-based approach. In: Proceedings of the 31st Meeting of the Association of Computational Linguistics. pp. 259–265.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics 21 (4), 543–565.

Brill, E., 1996. Learning to parse with transformations. In: Bunt, H., Tomita, M. (Eds.), Recent Advances in Parsing Technology. Kluwer Academic Press.

Cassier, L., Van de Craen, P., 1986. Vijftig jaar evolutie van het nederlands. In: Creten, J., Geerts, G., Jaspaert, K. (Eds.), werk-in-uitvoering. Leuven, Acco, pp. 59–74.

Cremelie, N., Martens, J., 1999. In search of better pronunciation models for speech recognition. Speech Communication 29 (2), 115–136.

Daelemans, W., van den Bosch, A., 1996. Language-independent data-oriented grapheme-to-phoneme conversion. In: Van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis. New York: Springer Verlag, pp. 77–90.

De Schutter, G., 1978. Aspekten van de Nederlandse klankstruktuur. Vol. 15. Antwerp Papers In Linguistics.

Freedman, D., Pisani, R., Purves, R., Adhikari, A., 1991. Statistics. W.W. Norton & Company.

Fukada, T., Yoshimura, T., Sagisaka, Y., 1999. Automatic generation of multiple pronunciations based on neural networks. Speech Communication 27 (1), 63–73.

Goossens, J., 1973. De belgische uitspraak van het nederlands. De nieuwe

taalgids 66 (3), 230–240.

Humphries, J., Woodland, P., Pierce, D., 1996. Using accent-specific pronunciation modelling for robust speech recognition. In: Proceedings of ICSLP. pp. 2324–2327.

Kessens, J., Wester, M., Strik, H., 1999. Improving the performance of a dutch csr by modeling within-word and cross-word pronunciation variation. Speech Communication 29 (2), 193–207.

Kipp, A., Wesenick, M.-B., Schiel, F., 1997. Pronunciation modeling applied to automatic segmentation of sponaneous speech. In: Proceedings of Eurospeech-97. pp. 1023–1026.

Livescu, K., 1999. Analysis and modeling of non-native speech for automatic speech recognition. Ph.D. thesis, MIT, Cambridge.

Quinlan, J., 1993. C4.5: programs for machine learning. San Mateo: Morgan kaufmann Publishers.

Ramshaw, L., Marcus, M., 1994. Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging. In: Proceedings of the Balancing Act Workshop on Combining Symbolic and Statistical Approaches to Language. pp. 86–95.

Ramshaw, L., Marcus, M., 1995. Text chunking using transformation based learning. In: Proceedings of the Third ACL Workshop on Very Large Corpora. pp. 82–94.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, M., 1999. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Communication 29 (2), 209–224.

Roche, E., Schabes, Y., 1995. Deterministic part-of-speech tagging with finite-state transducers. Computational Linguistics 21 (2), 227–253.

Sejnowski, T., Rosenberg, C., 1987. Parallel networks that learn to pronounce english text. Complex Systems 1, 145–168.

Smakman, D., van Bezooijen, R., 1999. De uitspraak van het standaardned-

erlands in nederland - een evaluatief en descriptief onderzoek. In: Huls, E., Weltens, B. (Eds.), Artikelen van de Derde Sociolingustische Conferentie. pp. 367–378.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for asr: a survey of the literature. Speech Communication 29 (2), 225–246.

Tomokiyo, L., Waibel, A., 2001. Adaptation methods for non-native speech. In: Proceedings of Multilinguality in Spoken Language Processing.

Van de Velde, H., 1996. Variatie en verandering in het gesproken standaard-nederlands (1935-1993). Ph.D. thesis, Katholieke Universiteit Nijmegen.

Van Haver, J., 1972. De uitspraak van het Nederlands. Leuven: Acco.

van Hout, R., De Schutter, G., De Crom, E., Huinck, W., Kloots, H., Van de Velde, H., 1999. De uitspraak van het standaard-nederlands. variatie en varianten in vlaanderen en nederland. In: Huls, E., Weltens, B. (Eds.), Artikelen van de Derde Sociolingustische Conferentie. pp. 183–196.

Wester, M., Fosler-Lussier, E., 2000. A comparison of data-derived and knowledge-based modeling of pronunciation variation. In: Proceedings of ICSLP 2000.

A Table with the phonetic alphabets IPA (I), DISC (D) and YAPA (Y)

Table A contains the phonemes present in Celex and Fonilex and their corresponding phonemic representation in IPA, DISC and YAPA. For some phonemes, DISC or YAPA do not have a particular phonemic representation. The <y> in "analyse" (Eng. "analysis"), for instance, is represented by the phonemic symbol /!/ in DISC. YAPA uses the same /i/ in "analyse" as in all words containing a close front unrounded vowel (e.g. "liep" (Eng. "ran")).

Word form		I	D	Y	Word form		I	D	Y
Dutch	English				Dutch	English			
put	well	p	p	p	**b**ad	bath	b	b	b
tak	branch	t	t	t	**d**ak	roof	d	d	d
kat	cat	k	k	k	**g**oal	goal	g	g	g
la**ng**	long	ŋ	N	N	**m**elk	milk	m	m	m
nat	wet	n	n	n	**l**at	slat	l	l	l
rat	rat	r	r	r	**f**iets	bycicle	f	f	f
vat	barrel	v	v	v	**s**ap	juice	s	s	s
zat	sat	z	z	z	**sj**aal	scarf	ʃ	S	S
rava**g**e	ravage	ʒ	Z	Z	**j**as	coat	j	j	j
gaat	goes	x	x	x	re**g**en	rain	ɣ	G	G
had	had	h	h	h	**w**at	what	ʋ	w	w
jazz	jazz	ʤ	-	dZ	li**e**p	ran	iː	i	i
r**uw**	rough	yː	y	y	l**ee**g	empty	eː	e	e
d**eu**k	dent	øː	|	&	l**aa**t	late	aː	a	a
b**oo**m	tree	oː	o	o	b**oe**k	book	uː	u	u
l**i**p	lip	ɪ	I	I	z**e**g	say	ɛ	E	E
l**a**t	slat	ɑ	A	A	b**o**m	bomb	ɔ	O	O
p**u**t	well	ʉ	}	Y	g**e**lijk	equal	ə	@	@
anal**y**se	analysis	iːː	!	i	centrif**u**ge	centrifuge	yːː	(	y

Word form		I	D	Y	Word form		I	D	Y
Dutch	English				Dutch	English			
sc**e**ne	scene	ɛː	)	E:	r**o**se	pink	ɔː	o	O:
co**ng**é	holiday	ɔ̃	On	O	vacc**i**n	vaccine	ɛ̃	)	E
croiss**ant**	croissant	ɑ̃	An	A	parf**um**	perfume	ɤ̃	}m	Y
fr**eu**le	lady	œː	*	@:	z**o**ne	zone	ɒː	<	o
w**ij**s	wise	ɛi	K	E:j	h**ui**s	house	œy	L	O:w
k**ou**d	cold	ɑu	M	@:9	ka**m**fer	camphor	ɱ	m	M
champa**gn**e	champagne	ɲ	nj	Jj					