Machine Learning for Modeling Dutch Pronunciation Variation.

Véronique Hoste, Steven Gillis, Walter Daelemans
CNTS - Language Technology Group, University of Antwerp

Monachesi (ed.), Proceedings 10th CLIN meeting, p. 73-84, 2000

Abstract

This paper describes the use of rule induction techniques for the automatic extraction of phonemic knowledge and rules from pairs of pronunciation lexicons. This extracted knowledge allows the adaptation of speech processing systems to regional variants of a language. As a case study, we apply the approach to Northern Dutch and Flemish (the variant of Dutch spoken in Flanders, a part of Belgium), based on Celex and Fonilex, pronunciation lexicons for Northern Dutch and Flemish, respectively. In our study, we compare two rule induction techniques, Transformation-Based Error-Driven Learning (TBEDL) (Brill, 1995) and C5.0 (Quinlan, 1993), and evaluate the extracted knowledge quantitatively (accuracy) and qualitatively (linguistic relevance of the rules). We conclude that, whereas classification-based rule induction with C5.0 is more accurate, the transformation rules learned with TBEDL can be more easily interpreted.

1 Introduction

A central component of speech processing systems is a pronunciation lexicon defining the relationship between the spelling and pronunciation of words. Regional variants of a language may differ considerably in their pronunciation. Once a speaker from a particular region is detected, speech input and output systems should be able to adapt their pronunciation lexicon to this regional variant. Regional pronunciation differences are mostly systematic and can be modeled using rules designed by experts. However, in this paper, we investigate the automation of this process by using data-driven techniques, or more specifically, rule induction techniques.

Data-driven methods have proven their efficacy in several similar language engineering tasks, such as grapheme-to-phoneme conversion, part-of-speech tagging, etc. Extraction of linguistic knowledge from a sample corpus instead of manual encoding of linguistic information proved to be an extremely powerful method for overcoming the linguistic knowledge acquisition bottleneck. Different approaches are available, such as decision-tree learning (Dietterich, 1997), neural network or connectionist approaches (Sejnowski and Rosenberg, 1987), memory-based learning (Daelemans and van den Bosch, 1996) etc. Data-driven approaches can yield results that are comparable to and often even better than rule-based approaches, as described in Daelemans and van den Bosch (1996) in which a comparison is made between Morpa-cum-Morphon (Nunn and van Heuven, 1993), an example of a linguistic knowledge based approach to grapheme-to-phoneme conversion and IG-Tree, an example of a memory-based approach (Daelemans et al., 1997).

In this study, we will look for the patterns and generalizations in the phonemic differences between Dutch and Flemish by using two data-driven techniques. It is our aim to extract the regularities that are implicitly contained in the data. Two corpora were used for this study, representing the Northern Dutch and Southern Dutch variants. For Northern Dutch Celex (release 2) was used and for Flemish Fonilex (version 1.0b). The Celex database contains frequency information as well as phonological, morphological, and syntactic information about more than 384.000 word forms. DISC is used as encoding scheme for word pronunciation. The Fonilex database is a list of more than 200.000 word forms with their Flemish pronunciation. For each word form, an abstract phonological representation is given, as well as the concrete pronunciation of that word form in three speech styles: highly formal speech, sloppy speech and "normal" speech (which is an intermediate level). A set of phonological rewrite rules was used to deduce these concrete speech styles from the abstract phonological form. The initial phonological transcription was obtained by a grapheme-to-phoneme converter and corrected by hand afterwards. Fonilex uses YAPA as encoding scheme. The Fonilex entries also contain a reference to the Celex entries, since Celex served as basis for the list of word forms in Fonilex. 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.

In the following section, a brief explanation is given of the method we used to find the overlap and differences between both regional variants of Dutch. Section 3 provides a quantitative analysis of the results. Section 4 discusses the differences between Celex and Fonilex, starting from the set of transformation rules that is learned during Transformation-Based Error-Driven Learning (TBEDL). These rules are compared to the production rules produced by C5.0. In addition, we present an overview of the non-systematic differences. In a final section, some concluding remarks are given.

2 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 overlap and discrepancies between both variants. Moreover, they can be used to adapt pronunciation databases for Dutch automatically to Flemish and vice versa. A possible way to find the regularities within the differences between both corpora is to make the rules by hand, which is time-consuming and error-prone. Another option is to make use of a data-oriented learning method in which linguistic knowledge is learned automatically. In our experiment we used two rule induction techniques, viz. Transformation-Based Error-Driven Learning (TBEDL) (Brill, 1995) and C5.0 (Quinlan, 1993).

2.1 TBEDL

In the process of Transformation-Based Error-Driven Learning, transformation rules are learned by comparing a corpus that is annotated by an initial state annotator to a correctly annotated corpus, which is called the "truth". During that comparison, an ordered list of transformation rules is learned. These rules are applied to the output of the initial state annotator in order to bring that output closer to the "truth". A rule consists of two parts: a transformation and a "triggering environment". For each iteration in the learning process, it is investigated for each possible rule how many mistakes can be corrected through application of that rule. The rule which causes the greatest error reduction is retained.

Figure 1 shows the TBEDL learning process applied to the comparison of the Celex-representation and the Fonilex "normal" representation. In the two TBEDL experiments that were performed, both variants function once as "truth". 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). This learning process results in an ordered list of transformation rules which reflects the systematic differences between both representations. A rule is read as: "change x (one representation) into y (other representation) in the following triggering environment".

E.g.: /iː/ /ɪ/ NEXT1OR2OR3PHON /eː/
(change a tense /iː/ to a lax /ɪ/ when one of the three following Celex phonemes is a tense /eː/)

To learn a transformation during the learning process, the learner applies every possible transformation, which means that all possible instantiations of the transformation templates are tried. These transformation templates specify a small number of features or feature sets that are relevant to finding an appropriate rule. In our task of deriving one variant of Dutch from the other variant, the graphemes and phonemes within a range of three positions to the left and three positions to the right of the target phoneme are used, e.g. "PREVPHON",

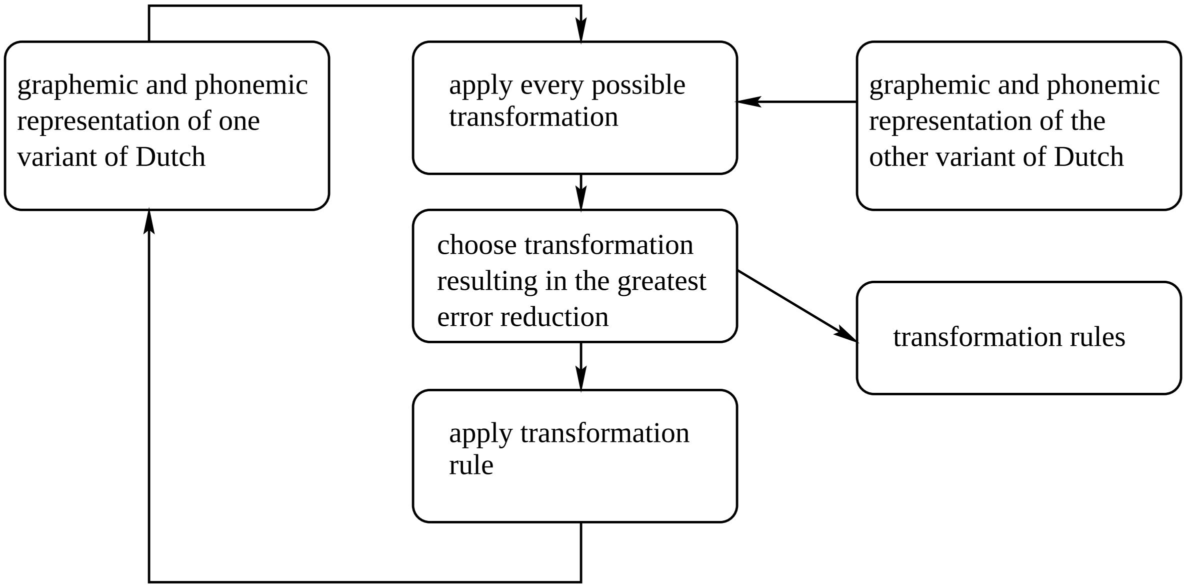

Figure 1: Architecture of the learning process making use of TBEDL

"NEXT1OR2GRAPH", "CURGRAPH", "LBIGRAM", etc. Rules also take into account word boundary information, which is indicated by "STAART". For each transformation application, the different transformation templates are applied to the cases where both corpora differ in phonemic representation. The transformation rule causing the greatest error reduction is chosen. In this experiment, the standard set of transformation templates provided in the Brill-learner is used, containing 26 different templates, as shown in Table 1. 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.

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

Table 1: Set of transformation templates used in the learning process

2.2 C5.0

C5.0 (Quinlan, 1993), which is a commercial version of the C4.5 program, generates a classifier in the form of a decision tree. This decision tree classifies a case starting at the root of the tree and then moving through the tree until a leaf node (associated with a class) is encountered. Since decision trees for this application can be hard to read, the decision tree is converted to a set of production rules, which are more intelligible to the user. The rules have the form "L -> R", in which the left-hand side is a conjunction of attribute-based tests and the right-hand side is a class. When classifying a 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 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 $\{V_1, V_2, ...\}$". The method Quinlan (1993) uses to find groups of attribute values, is based on iterative merging of value groups. The partitioning of the training cases is based on the *gain ratio criterion*, which expresses the amount of information generated by the split of the training cases that appears helpful for classification. This grouping based on statistical information sometimes makes it hard to understand the production rules, because the value groups are not always a reflection of what is called in phonological theory "a natural class", which is a coherent grouping of phonetically similar sounds.

The input pattern in our experiment consists of graphemic and phonemic information. The task is defined as the conversion of fixed-size instances representing the focus grapheme ('fg') and phoneme ('fp'), with a certain context to a class representing the target phoneme, as shown in Table 2, using a windowing technique proposed by Sejnowski and Rosenberg (1987).

Table 2: Example of instances generated from the word "kraker" (Eng. "squatter") for the C5.0 experiment.

graphemic representation			phonemic representation			class
left	**fg**	right	left	**fp**	right	
===	k	rak	===	k	raːk	k
==k	r	ake	==k	r	aːkə	r
=kr	a	ker	=kr	aː	kər	aː
kra	k	er=	kraː	k	ər=	k
rak	e	r==	raːk	ə	r==	ə
ake	r	===	aːkə	r	===	r

In the experiment, 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 predicted class for this case is then the right-hand side of the rule. At the top of the rule the number of training cases covered by the rule is given together with the number of covered cases that do not belong to the class predicted by the rule. The "lift" is the estimated accuracy of the rule divided by the prior probability of the predicted class.

E.g.: (1072/4, lift 724.2)
fg in {a, A, g, j, e, t, n, i, d, k, l, b, r, u, w, m, o, z, p, h, v, f, y, q, x, D, J, E, F, B, C, M, K, G, H, I, L, O, N, S, V, R, P, Q, T, U, W, X, Y, Z}
fp-1 in {aː, eː, iː, oː, yː}
fp = s
fp+1 in {j, ʋ, m, iː, juː, ɪj, dz, ɑj, aːj}
-> class ts [0.995]

2.3 Alignment

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/. This problem is solved by defining a new phonemic symbol that corresponds to the two phonemes, as indicated in Table 3.

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

Table 3: The use of compounds in "taxi".

Furthermore, alignment is required (Daelemans and van den Bosch, 1996), since the phonemic representation and the spelling of a word often differ in length. 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 4. In this experiment, alignment was performed for the graphemic and phonemic representations of Celex and for those of Fonilex.

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

Table 4: Alignment of the word "aalmoezenier" (Eng.: "chaplain").

The dataset we used consists of all Fonilex entries with omission of the double transcriptions, which represent ca. 1/20 of the corpus. In this case, only the first transcription is taken, as in the word “caravan”, which can be phonemically represented as /kɑrɑvɑn/ or as /kɛrɛvɛn/. Words of which the phonemic transcription is longer than the orthography and for which no compound phonemes are provided, are omitted, e.g. ”b’tje” (Eng.: “little b”)(phonemically: /beːtjə/).

Both the use of compound phonemes and alignment lead to a corpus consisting of 202.136 records or 1.972.577 phonemes. DISC is used as phonemic encoding scheme. All DISC phonemes are included and new phonemes are created for the phonemic symbols which only occur in the Fonilex database. We have divided the corpus into a training part, consisting of 90% of the data and a 10% test part.

Initially, an overlap of 59.07% on the word level and 92.77% on the phoneme level was observed in the 10% test set between the Dutch and Flemish representations. Consonants and diphthongs are highly overlapping.

Word	Phon.	Cons.	Vowel	Diph.
59.07	92.77	95.95	85.58	99.76

Table 5: Initial overlap between Celex en Fonilex

3 Quantitative analysis

We first test whether rule induction techniques can learn to adapt Northern Dutch pronunciations to Flemish when trained on a number of examples and vice versa. By using Transformation-Based Error-Driven Learning and C5.0, we looked for the systematic differences between Northern Dutch and Flemish.

In TBEDL, the complete training set of 90% was used for learning the transformation rules. A threshold of 15 errors was specified, which means that learning stops if the error reduction lies under that threshold. Due to the large amount of training data, this threshold was chosen to reduce training time. This resulted in ca. 450 transformation rules for the conversion of Celex into Fonilex and into ca. 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 Northern Dutch corpus. E.g. in “grandeur” (Eng.: “splendor”), the <n> is represented as /˜/ in Fonilex and as /n/ in Celex.

In Figure 2, the number of transformation rules is plotted against the accuracy of the conversion between Celex and Fonilex. A first comparison between both plots clearly shows the same tendencies in the accuracy percentages both on the word and the phoneme level. This figure indicates that, for both deriving Celex transcriptions from Fonilex transcriptions and vice versa, 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 applying 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. For the derivation of Northern Dutch pronunciation, accuracy increases from 83.0% to 88.2% (words) and from 97.6% to 98.5% (phonemes).

For the C5.0 experiment, 50% (887.647 cases) of the original training set served as training set (more training data was not feasible). A decision tree model and a production rule model were built from the training cases. The tree was converted to a set of 709 rules for the conversion of Celex transcriptions into Fonilex transcriptions. When learning Celex pronunciation, 658 rules were learned. These production rules were applied to the original 10% test set we used in the Brill experiment. In order to make the type of task comparable for the transformation based approach used by TBEDL and the classification-based approach used in C5.0, the output class to be predicted by C5.0 was either

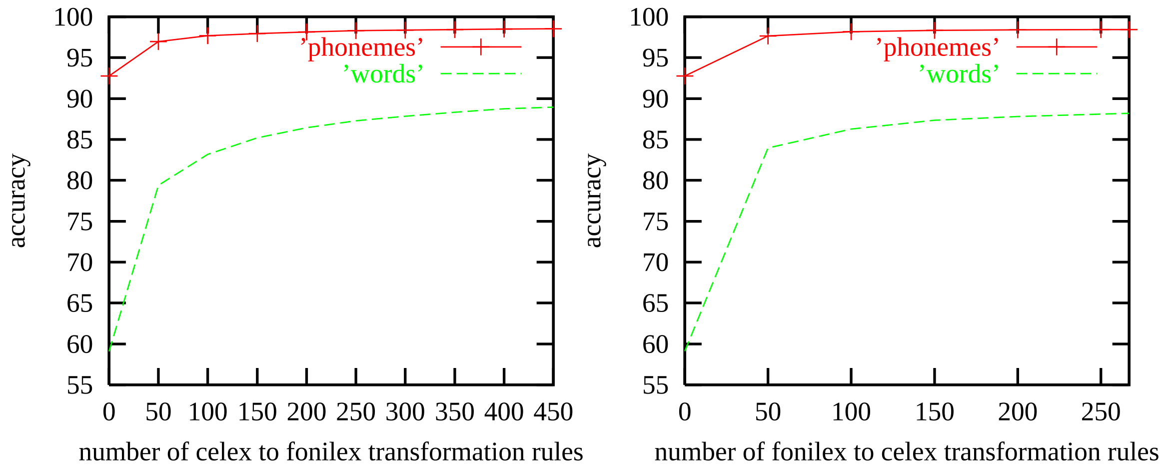

Figure 2: Description of the accuracy of the word and phoneme level in relation to the number of transformation rules

'0' when the Celex and Fonilex phoneme are identical (i.e. no change), or the target phoneme when Celex and Fonilex differ. 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 does, 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 test set, the transformation rules learned by the Brill-tagger have a higher error rate, even when C5.0 is only trained on half the data used by TBEDL.

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. On the phoneme level, the use of the Brill-tagger leads to a 98.5% accuracy. With a 98.9% accuracy, C5.0 outperforms the Brill-tagger.

When learning the Northern Dutch pronunciation, the same tendency can be observed. After application of the transformation rules, there is an 88.2% accuracy on the word level. When applying all C5.0 rules, 92.9% of the words are equally pronounced in Northern Dutch and Flemish. With regard to the overlapping phonemes, a 98.5% accuracy is observed when using TBEDL and a 99.1% when using C5.0.

In both learning experiments, C5.0 also has a slightly lower error rate for the consonants, vowels and diphthongs.

A comparison of the initial overlap between both variants 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 more or less 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).

It is indeed possible to reliably 'translate' Dutch into Flemish and vice versa.

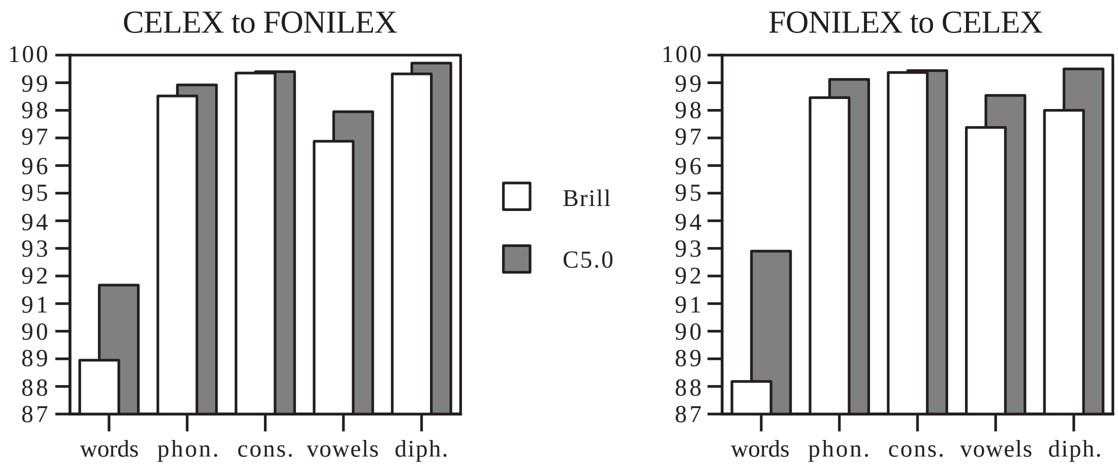

Figure 3: Accuracy after application of all transformation rules and C5.0 production rules

Nr.	CE	FO	Triggering environment
1.	x	ɣ	PREV 1 OR 2 PHON STAART
2.	iː	ɪ	NEXT 1 OR 2 OR 3 PHON eː
3.	j	tʃ	SURROUND PHON tə
4.	t	-	NEXT PHON tʃ
5.	iː	ɪ	NEXT 1 OR 2 GRAPH c
6.	iːj	ɪj	CUR GRAPH i
7.	oː	ɔ	NEXT 1 OR 2 OR 3 PHON eː
8.	ts	s	RBIGRAM t i
9.	aː	ɑ	NEXT 2 GRAPH a
10.	ʋ	-	PREV PHON ɑu

Nr.	FO	CE	Triggering environment
1.	ɣ	x	PREVPHON STAART
2.	ɪ	iː	NEXT 1 OR 2 GRAPH e
3.	tʃ	j	NEXT 1 OR 2 OR 3 PHON ə
4.	-	t	NEXT BIGRAM jə
5.	ɑ	aː	NEXT 1 OR 2 GRAPH i
6.	ɔ	oː	NEXT 1 OR 2 GRAPH e
7.	ɪ	iː	NEXT 2 GRAPH i
8.	ɔ	oː	NEXT 2 GRAPH i
9.	ɪj	iːj	CUR GRAPH i
10.	ɑ	aː	GRAPH AND 2 AFT a e

Table 6: Overview of the first ten rules learned during TBEDL. In the table on the left the Celex phonemes are converted into Fonilex phonemes. In the table on the right, the rules of the conversion in the opposite direction are given.

4 Qualitative Analysis

In this section, we are interested in the linguistic quality of the rules that were extracted using TBEDL and C5.0. To gain more insight in the important differences between the two pronunciation variants, a qualitative analysis of the rules was performed. Therefore, the conversion rules were listed and compared. The following list presents some examples for consonants, vowels and diphthongs. We will discuss the first 10 rules that were learned during TBEDL, as shown in Table 6, which will be compared with the 10 non-0 production rules, which most reduce the error rate.

The transformation rules presented in Table 6, formulate the most important pronunciation differences between Northern Dutch and Flemish in a set of easily understandable rules. The C5.0 production rules, on the other hand also describe the overlapping phonemes between Celex and Fonilex, which makes it hard to have a clear overview of the regularities in the differences between both variants of Dutch. The fact that the category '0' was used to describe the overlap between the databases (no change) does not really help. Even if C5.0 discovers that no change is the default rule, additional specific rules describing the default condition are nevertheless necessary to prevent the other rules from firing incorrectly. Another disadvantage of the C5.0 rules is that, in our opinion, these production rules are harder to interpret than the Brill-rules due to the value grouping mechanism, described in section 2.2., which can lead to groupings in which feature values do not necessarily correspond to phonological reality. A comparison of the second transformation rule in the learning process of Northern Dutch pronunciation (see Table 6) and the following C5.0 rule clearly illustrates this phenomenon:[1]

(8717/111, lift 87.3)
fg-1 in {a, g, j, t, e, i, d, n, l, b, s, r, u, k, w, m, o, z, p, c, h, v, f, y, x, J, F, B, C, M, H, L, N, S, P, T, W}
fg+2 in {e, i, u, a, o, c, y}
fp = ɪ
fp+1 in {t, d, n, s, k, l, b, ʐ, r, ɣ, ʋ, m, z, p, v, h, f, iːj, ɡ, dz}
fp+2 in {j, -, ɛ, œy, eː, ɑ, ʋ, uː, iː, ʉ, ɔ, aː, ɪ, oː, yː, iːj, Y9, ɪj, ɛj, aːj, ɒː, ɛː, eːj, oʊ, ɔʊ, əː, ɑj, y9}
fp+3 in {=, j, ə, t, -, d, n, s, b, ɛ, ʐ, l, ɛi, r, œy, k, ɣ, eː, ɑ, ʋ, m, uː, z, iː, p, ʉ, x, ɔ, v, h, aː, ɪ, ŋ, ks, f, oː, ts, yː, ɑu, yːʋ, œː, @?, øː, Jj, juː, ɡ, Y9, ɪj, ˜, dʐ, ɛj, aːj, ɡz, ɒː, iːj, ɛː, dz, eːj, oːʋ, ɔʊ, əː, ɑj, uːʋ, juːʋ, y9}
-> class iː[0.987]

4.1 Consonants

When looking at the differences on the consonant level, nearly 60% of the differences on the consonant level 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 an /s/ in /mɑɣiːjaːvɛlɪsmə/ in Dutch and with a /z/ in /mɑkɪjɑvɛlɪzmə/ in Flemish.

A closer look at the confusion matrix in Table 7 shows that especially the alternation between /x/ and /ɣ/ is very frequent. This alternation

[1] In those cases where no IPA-equivalent exists for the phonemes mentioned in this rule, the DISC-phonemes are used. If no DISC-phoneme is available, the YAPA-phonemes are used. The compound phonemes are also converted back into the original phonemic combinations.

	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

Table 7: Confusion matrix for the voiced and unvoiced consonants in the test corpus.

is also the subject of the first transformation rule that was learned in both directions of the conversion process, namely "/x/ changes into /ɣ/ in case of a word boundary one or two positions before" when converting the Celex pronunciation into the Fonilex pronunciation. For the conversion of the Flemish Fonilex pronunciation into the Northern Dutch Celex pronunciation, the rule "/ɣ/ changes into /x/ in case of a word boundary one position before" is learned. When looking at the top ten of the C5.0 production rules that most reduce error rate, two important rules also describe this alternation:

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

Fonilex to Celex :
(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, ʐ, ɛi, r, (...)}
-> class x [0.993]

Another important phenomenon is the use of palatalization in Flemish, as in the word "aaitje" (Eng.: "stroke"), where Fonilex uses the palatalized form /aːjtʃə/ instead of /aːjtjə/. The two subsequent Brill rules 3 and 4 (see Table 6) make this change possible. When learning Flemish pronunciation, the /j/ is first changed into /tʃ/in case of the surrounding phonemes /t/ and /ə/. In rule 4, the Dutch /t/ is omitted if the following phoneme is a /tʃ/. When learning the Dutch pronunciation of the diminutive ending “tje”, the same is learned but in the opposite direction. As a first step, /tʃ/ changes into /j/. In a second step, a /t/ is added in front of the bigram /jə/. This change in both directions is also described in the top 10 of C5.0 rules.

4.2 Vowels

96% of the differences at the vowel level between Dutch and Flemish concerns the use of a lax vowel instead of a tense vowel for the /iː/, /eː/, /aː/, /oː/ en /uː/. This alternation is illustrated by the following confusion matrix, which clearly shows that tense Celex-vowels not only correspond with tense, but also with lax vowels in Fonilex. Other less frequent differences are glide insertion, e.g. in “geshaket” and the use of schwa instead of another vowel, as in “teleprocessing” in Flemish.

	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

Table 8: Confusion matrix showing the use of Flemish lax and tense vowels given the Dutch tense vowels.

For the conversion of the Northern Dutch pronunciation to the Flemish pronunciation, the transformation rules 2, 5, 6, 7 and 9, as shown in Table 6, describe the transition from a tense vowel into a lax vowel in a certain triggering environment. An example is the word “multipliceer” (Eng.: “multiply”) which is transcribed as /mʉltiːpliːseːr/ in Celex and as /mʉltɪplɪseːr/ in Fonilex.

When learning the pronunciation of Northern Dutch vowels, the transition from lax vowels (such as /ɪ/, /ɑ/, /ɔ/) into the corresponding tense vowels (/iː/, /aː/, /oː/) is clearly shown in the first ten rules (see transformation rules 2, 5, 6, 7, 8, 9, 10).

A closer look at the ten most important C5.0 production rules shows that for both learning Northern Dutch and Flemish pronunciation, seven out of ten rules describe this alternation between a tense and a lax vowel. E.g.

Celex to Fonilex:
(4370/138, lift 82.8)
fp = iː
fp+2 in {ɛ, ʐ, eː, aː, yː, ʃ, ɛː}
-> class ɪ [0.968]

Fonilex to Celex :
(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]

4.3 Diphthongs

For the diphthongs, few transformation rules are learned during training, since Celex and Fonilex are highly overlapping (see Table 1). The rules concern the phonemes that follow the diphthongs: /j/ after /ɛi/ and /ʊ/ after /ɑu/. E.g. in “blauw” (Eng.: “blue”), the /ʊ/ is omitted in Flemish: /blɑu/. Learning Flemish pronunciation gave rise to the following top ten rule: “/ʊ/ is omitted if the preceding phoneme is an /ɑu/”. In the other TBEDL experiment and in both C5.0 experiments, no top ten rules describing the lack or presence of /j/ or /ʊ/ after diphthongs, were given.

These rules, describing the differences between Northern Dutch and Flemish consonants, vowels and diphthongs also make linguistic sense. Linguistic literature, such as Booij (1995) 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 Northern Dutch and Flemish. The same discrepancies are found in the transcriptions made by Flemish subjects in the transcription experiments described in Gillis (1999). In this experiment, a comparison of an example transcription and the transcription made by different persons reveals that the important differences between Northern Dutch and Flemish, namely the alternations between voiced and unvoiced consonants and the tendency to use lax vowels in Flemish and tense vowels in Northern Dutch lead to confusion in the transcription choices. The largest part of the differences from the example transcription can be reduced to a limited number of substitutions. The most important substitution patterns on the vowel level concern the substitution of a tense vowel by its lax counterpart and vice versa. On the consonant level, a voiced obstruent is often substituted by its unvoiced counterpart.

5 Error Analysis

Besides the systematic phonemic differences between Flemish and Dutch, there are a number of unsystematic differences between both databases. After application of the transformation rules, 89.0% of the words makes a correct transition from the Celex -transcription to the Fonilex-transcription and 88.2% of the words makes the correct transition in the opposite direction. The C5.0 rules lead to a 91.7%, when learning the Flemish pronunciation and a 92.9%, when learning the Northern Dutch pronunciation.

Using the Brill-tagger, it has also to be taken into account that 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, which does not suffer from similar consequences of rule ordering, will correctly classify "feuilleteer".

In this section, we are concerned with the remaining errors after application of all rules. In this error analysis, the conversion of Northern Dutch into Flemish was studied. Making use of a rule induction technique to extract the sub-regularities in the differences between the corpora can lead to some rules, which, however, may be based on noise or errors in the databases. Therefore, a manual analysis was done, which 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 big 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 are loan words, such as "points" and "panty's" or the loan sound /˜/ which only appears in Fonilex.

Another cause for errors is that rules will overgeneralize in certain cases. The confusion matrix for vowels in Table 8 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. A closer investigation of the errors committed by the Brill-tagger, however, 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.

6 Concluding remarks

In this paper, we have proposed the use of rule induction techniques to learn to adapt pronunciation representations to regional variants, and to study the linguistic aspects of such variation. A quantitative and qualitative analysis was given of the phonemic differences discovered by these techniques when trained on the Celex database (Dutch) and the Fonilex database (Flemish). In order to study the relationship between both pronunciation systems, we used two rule induction techniques, namely Transformation-Based Error-Driven Learning (Brill, 1995) and C5.0 (Quinlan, 1993).

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) was obtained. For the conversion of Flemish into Northern Dutch pronunciation, the same tendencies can be observed: an overall accuracy of 88% is reached in predicting the pronunciation of a northern Dutch word when applying the transformation rules. When applying all C5.0 rules, 93% of the words are equally pronounced in Northern Dutch and Flemish. With respect to the phonemes, a 98% accuracy is observed when using TBEDL and a 99% when using C5.0. The C5.0 production rules prove to be more accurate in predicting Northern Dutch and Flemish pronunciation.

The accuracies of both learning techniques indicate that it is indeed possible to reliably convert Northern Dutch into Flemish and vice versa. Moreover, the use of these rule-induction techniques can be an appropriate method for adapting pronunciation databases of one variant automatically to the other variant.

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 C5.0 production rules, however, are more numerous and more complex than the transformation rules. Furthermore, the C5.0 rules also describe the overlapping phonemes in both variants of Dutch, which makes it hard to have a clear overview of the regularities in the differences between Flemish and Northern Dutch. The results of the transformation-based learning approach are clearly more understandable than those of a classification-based learning approach for this problem.

Acknowledgements

Part of the research was published earlier as (Hoste et al., 2000). This research was partially funded by the FWO project Linguaduct and the IWT project CGN (Corpus Gesproken Nederlands).

References

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

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

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

W. Daelemans, A. van den Bosch, and T. Weijters. 1997. Igtree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11:407–423.

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

T.G. Dietterich. 1997. Machine learning research: Four current directions. *AI Magazine*, 18(4):97–136.

S. Gillis. 1999. Phonemic transcriptions: qualitative and quantitative aspects. Paper presented at the International Workshop about Design and Annotation of Speech Corpora, Tilburg.

V. Hoste, G. Gillis, and W. Daelemans. 2000. A rule induction approach to modeling regional pronunciation variation. In *COLING 2000 - Proceedings of the 18th International Conference on Computational Linguistics*.

A. Nunn and V.J. van Heuven. 1993. Morphon, lexicon-based text-to-phoneme conversion and phonological rules. In V.J. Van Heuven and L.C.W. Pols, editors, *Analysis and synthesis of speech; strategic research towards high-quality text-to-speech generation*. Berlin, Mouton de Gruyter.

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

L.A. Ramshaw and M.P. Marcus. 1995. Text chunking using transformation based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94.

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

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