A Rule Induction Approach to Modeling Regional Pronunciation Variation.

Véronique Hoste and **Steven Gillis** and **Walter Daelemans***
CNTS - Language Technology Group
University of Antwerp
Universiteitsplein1, 2610 Wilrijk
hoste@uia.ua.ac.be, gillis@uia.ua.ac.be, daelem@uia.ua.ac.be

Proceedings 18th international conference on computational linguistics
Saarbruecken, p. 327-333, 2000

Abstract

This paper describes the use of rule induction techniques for the automatic extraction of phonemic knowledge and rules from pairs of pronunciation lexica. 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 lexica 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, more specifically, rule induction techniques.

Data-driven methods have proven their efficacy in several 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 comparable (and often even better) results than the rule-based approach, as described in the work of Daelemans and van den Bosch (1996) in which a comparison is made between Morpa-cum-Morphon (Heemskerk 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., 1996).

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 (based on the INL corpus of the Institute for Dutch Lexicology), and phonological, morphological, and syntactic lexical information for more than 384.000 word forms,

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

and uses DISC as encoding for word pronunciation. The Fonilex database is a list of more than 200.000 word forms together with their Flemish pronunciation. For each word form, an abstract lexical representation is given, together with 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 was afterwards corrected by hand. Fonilex uses YAPA as encoding scheme. By means of their identification number, the Fonilex entries also contain a reference to the Celex entries, since Celex served as basis for the list of word forms in Fonilex. E.g. for the word "aaitje" (Eng.: "stroke"), the relevant Celex entry is "25/aaitje/5/'aj-tj@/" and the corresponding Fonilex entry looks like "25|aaitje|'ajtS@|". The word forms in Celex with a frequency of 1 and higher (indicated in field 3) are included in Fonilex and from the list with frequency 0, only the monomorphematic words were selected.

In the following section, a brief explanation is given of the method we used to search for 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. Hence, 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 have made use of two rule induction techniques, viz. Transformation-Based Error Driven Learning (TBEDL) (Brill, 1995) and C5.0 (Quinlan, 1993).

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. This ordering implies that the application of an earlier rule sometimes makes it possible for a later rule to apply (so-called "feeding"). In other cases, as also described in the work of Roche and Schabes (1995), a given structure fails to undergo a rule as a consequence of some earlier rule ("bleeding"). 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, which functions as "truth". In this case, the task is to learn how to transform Celex representations into Fonilex representations (i.e., translate Dutch pronunciation to Flemish pronunciation). 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 (Celex representation) into y (Fonilex representation) in the following triggering environment".
E.g. /iː/ /ɪ/ NEXT 1 OR 2 OR 3 PHON /eː/ (change a tense /i/ to a lax /i/ when one of the three following Celex phonemes is a tense /e/).

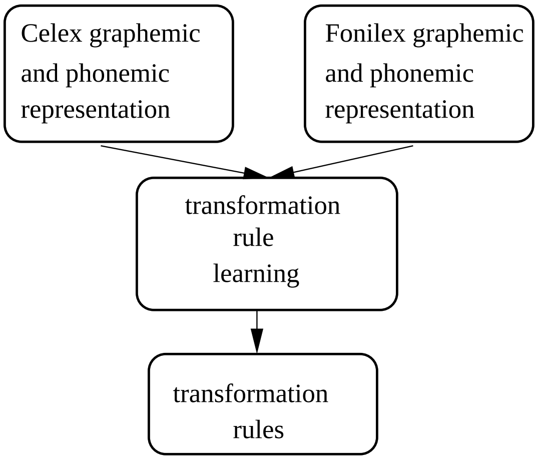

Figure 1: Architecture of the learning process making use of TBEDL

C5.0 (Quinlan, 1993), on the other hand, which is a commercial version of the C4.5 program, generates a classifier in the form of a decision tree. This decision tree can be used to classify a case by 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 can be hard to read, the decision tree is converted to a set of production rules, which are more intelligible to the user. All 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. Note that in the implementation of C5.0, feeding and bleeding effects of rules do not occur, due to the conflict resolution strategy used, which ensures that for each case only one rule can apply (Quinlan, 1993). In this experiment we have made use of a context of three phonemes preceding (indicated by f-1, f-2, and f-3) and following (f+1, f+2, f+3) the focus phoneme, which is indicated by an ’f’. 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 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.: (4370/138, lift 82.8)
f = iː
f+2 in {ɛ, ʐ, eː, aː, yː, ʃ, ɛː}
-> class ɪ [0.968]

Before presenting the data to TBEDL and C5.0, alignment is required (Daelemans and van den Bosch, 1996) for the graphemic and phonemic representations of Celex and Fonilex, 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. In the example “aalmoezenier” (Eng.: “chaplain”) this results in:

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

A further step in the preparation of the data, consists of the use of an extensive set of so-called “compound phonemes”. 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.

Our dataset consists of all Fonilex entries with omission of the double transcriptions (only the first transcription is taken), as in the word “caravan”, which can be phonemically represented as /kɑrɑvɑn/ or as /kɛrɛvɛn/. Also 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/). The corpus consists of 202.136 word forms 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 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 1: Initial overlap between Celex en Fonilex

3 Quantitative analysis

We first test whether rule induction techniques are able to learn to adapt Northern Dutch pronunciations to Flemish when trained on a number of examples. With 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 about 450 rules. In figure 2, the number of transformation rules is plotted against the accuracy of the conversion between both variants.

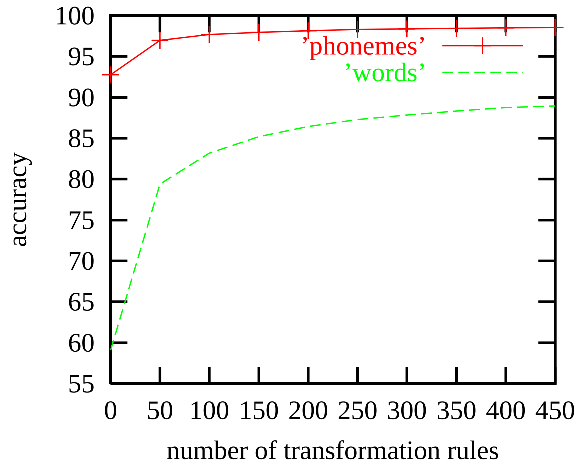

Figure 2: Description of the accuracy of the word and phoneme level in relation to the number of transformation rules.

Figure 2 shows that especially the first 50 rules lead to a considerable increase of performance from 59.07% to 79.40% on the word level and from 92.77% to 96.98% for phonemes, which indicates the high applicability of these rules. Afterwards, the increase of accuracy is more gradual: from 79.40% to 88.95% (words) and from 96.98% to 98.52% (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 gave rise to 745 rules. These production rules were applied to the original 10% test set we used in the TBEDL 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 '0' when the Celex and Fonilex phoneme are identical (i.e. no change), or the Fonilex phoneme when Celex and Fonilex differ.

Table 2 gives an overview of the overlap between Celex and Fonilex after application of both rule induction techniques. A comparison of these results shows that, when evaluating both TBEDL and C5.0 on the test set, the 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. On the word level, the initial overlap of 59.07% is raised to 88.95% after application of the 450 transformation rules, and to 90.35% when using the C5.0 rules. On the phoneme level, the initial 92.77% overlap is increased to 98.52% (TBEDL) and 98.74% (C5.0). C5.0 also has a slightly lower error rate for the consonants, vowels and diphthongs.

	Word	Phon.	Cons.	Vowel	Diph.
Brill	88.95	98.52	99.35	96.88	99.32
C5.0	90.35	98.74	99.19	97.70	99.68

Table 2: Overlap between Celex en Fonilex after application of 450 transformation rules and all C5.0 production rules.

When looking at those cases where Celex and Fonilex differ, we see that it is possible to learn Brill 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 3% more accurate: 76.4% (words) and 82.6% (phonemes). It is indeed possible to reliably 'translate' Dutch into Flemish.

4 Qualitative Analysis

In this section, we are concerned with the linguistic quality of the rules that were extracted using TBEDL and C5.0. To gain more insight in the important differences between both 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. Starting point is the first 10 rules that were learned during TBEDL, which are compared with the 10 C5.0 rules, which most reduce the error rate. In the transformation rules presented below, the relationship between Dutch and Flemish, especially the most important differences, are extracted from the corpora and formulated 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.

4.1 Consonants

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 a /xəlɛikaːrdəx/ with a voiceless velar fricative in Dutch and /ɣəlɛikaːrdəx/ with a voiced velar fricative in Flemish. The word "machiavellisme" (Eng.: "Machiavellism") is pronounced as /mɑɣiːjaːvɛlɪsmə/ in Dutch and as /mɑkɪjɑvɛlɪzmə/ in Flemish.

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

Table 3 clearly shows the alternation between /x/ and /ɣ/. This alternation also is the subject of the first transformation rule, namely "/x/ changes into /ɣ/ in case of a word beginning (indicated by "STAART") one or two positions before". When looking at the top ten of the C5.0 production rules that most reduce error rate, the two most important rules also describe this alternation:

Rule 682: (7749/29, lift 110.9)
f-1 in {=, {, ɛː}
f in {x, g, ;, Q}[1]
-> class ɣ [0.996]

Rule 683: (7749/29, lift 110.9)
f-1 in {=, {, ɛː}
f in {x, g}
-> class ɣ [0.996]

Another important phenomenon is the use of palatalisation in Flemish, as in the word "aaitje" (Eng.: "stroke"), where Fonilex uses the palatalized form /aːjtʃə/ instead of /aːjtjə/. The two subsequent transformation rules 3 and 4 make this change possible. In the top 10 of C5.0 rules, only the first part of this change is described. Transformation rule 8 describes the omission of the phoneme /t/ in case of the graphemic combination <ti>, as in "politie" (Eng.: "police").

Nr.	C.	F.	Triggering environment
1.	x	ɣ	PREV 1 OR 2 PHON STAART
3.	j	tʃ	SURROUND PHON tə
4.	t	-	NEXT PHON tʃ
8.	ts	s	RBIGRAM t i

Table 4: Transformation rules for the most frequent differences at the consonant level.

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

[1] The /;/ and /Q/ are compound phonemes we introduced. They do not have an IPA equivalent.

In transformation rules 2, 5, 6, 7, 9, there is a 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.

Nr.	C.	F.	Triggering environment
2.	iː	ɪ	NEXT 1 OR 2 OR 3 PHON eː
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
9.	aː	ɑ	NEXT 2 GRAPH a

Table 6: Most important transformation rules for the differences between the Dutch and Flemish vowels.

A closer look at the ten most important C5.0 production rules shows that seven out of ten rules describe this transition from a Celex tense vowel to a Fonilex lax vowel. E.g.

Rule 322: (4370/138, lift 82.8)
f = iː
f+2 in {ɛ, ʐ, eː, aː, yː, ʃ, ɛː}
-> ɪ [0.968]

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 /ʋ/ after /ɑu/. E.g. in "blauw" (Eng.: "blue"), the /ʋ/ is omitted in Flemish: /blɑu/. In the top ten of C5.0 rules, no rules are given describing this phenomenon.

Nr.	C.	F.	Triggering environment
10.	ʋ	-	PREV PHON ɑu

Table 7: Transformation rule concerning the lack or presence of a /ʋ/ following an /ɑu/.

These rules, describing the differences between Northern Dutch and Flemish consonants, vowels and diphthongs also make linguistic sense. Linguistic literature, such as the work of 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 both Flemish and Dutch subjects in the Dutch transcription experiments described in Gillis (1999).

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 450 transformation rules, 88.95% of the words makes a correct transition from the Celex-transcription to the Fonilex-transcription. The 745 C5.0 rules lead to a 90.35%. Using the Brill-tagger, it also has 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. 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 "pantys" or the loan sound /˜/ which only appears in Fonilex.

Another cause for errors is that rules will overgeneralise in certain cases. The confusion matrix for vowels in table 5 clearly indicates the tendency to use more lax vowels in Flemish. This leads to a number of transformation rules 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% 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 analyse 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 have made use of two rule induction techniques, namely Transformation-Based Error-Driven Learning (Brill, 1995) and C5.0 (Quinlan, 1993). Studying the accuracy of both systems, we noted that after application of the transformation rules that were learned by the TBEDL method, 73% of the differences on the word level and 80% of the differences on the phoneme level was covered by the rules. The C5.0 percentages are some 3% higher. This corresponds with an overall accuracy in predicting the pronunciation of a Flemish word pronunciation from the Dutch pronunciation of about 89% for TBEDL and 90% for C5.0 (about 99% at phoneme level for both).

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 difficult to interpret. The results of the transformation-based learning approach are clearly more understandable than those of a classification-based learning approach for this problem.

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*, 21:543–565.

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

W. Daelemans, A. van den Bosch, and T. Weijters. 1996. Igtree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review, special issue on Lazy Learning.*

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.

J. Heemskerk and V.J. van Heuven. 1993. *MORPHA, a lexicon-based MORphological PArser.* Berlin, Mouton de Gruyter.

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

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.