The Acquisition of Stress: A Data-Oriented Approach

Walter Daelemans*
Tilburg University

Steven Gillis†
National Fund For Scientific Research, Belgium

Gert Durieux‡
University of Antwerp

A data-oriented (empiricist) alternative to the currently pervasive (nativist) Principles and Parameters approach to the acquisition of stress assignment is investigated. A similarity-based algorithm, viz. an augmented version of Instance-Based Learning is used to learn the system of main stress assignment in Dutch. In this nontrivial task a comprehensive lexicon of Dutch monomorphemes is used instead of the idealized and highly simplified description of the empirical data used in previous approaches.

It is demonstrated that a similarity-based learning method is effective in learning the complex stress system of Dutch. The task is accomplished without the a priori knowledge assumed to pre-exist in the learner in a Principles and Parameters framework.

A comparison of the system's behavior with a consensus linguistic analysis (in the framework of Metrical Phonology) shows that ease of learning correlates with decreasing degrees of markedness of metrical phenomena. It is also shown that the learning algorithm captures subregularities within the stress system of Dutch that cannot be described without going beyond some of the theoretical assumptions of metrical phonology.

1. Introduction

1.1 Metrical Phenomena and Theory

Machine learning of metrical phenomena is an interesting domain for exploring the potential of particular machine learning techniques. First of all, the assignment of stress in monomorphemic words, the subject of this paper, has been fairly well studied in metrical phonology. Within this framework, the stress patterns of numerous languages have been described in considerable detail. Thus, a solid theoretical framework as well as elaborate descriptions of the linguistic data are available. Moreover, learning metrical phenomena has been cast in terms of the Principles and Parameters approach (Chomsky 1981), which provides both the basic parameters along which possible stress systems may vary, and makes strong claims about the allegedly innate knowledge of the natural language learner.

* Institute for Language Technology and AI (ITK), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: walter.daelemans@kub.nl
† Center for Dutch Language and Speech, University of Antwerp, UIA, Universiteitsplein 1, 2610 Wilrijk, Belgium. E-mail: steven.gillis@uia.ac.be
‡ Center for Dutch Language and Speech, University of Antwerp, UIA, Universiteitsplein 1, 2610 Wilrijk, Belgium. E-mail: gert.durieux@uia.ac.be

Secondly, the domain of metrical phenomena can be studied as a (relatively) independent problem domain (unlike other domains such as, for instance, linguistic pragmatics, that typically have multiple dependencies with other domains like syntactic and/or semantic phenomena).

Thirdly, metrical phenomena exhibit a number of interesting characteristics that make them well suited for testing the capacity of machine learning algorithms to generalize as well as handle irregularities. On the one hand, stress assignment appears to be governed by a number of solid generalizations. For instance, we found that in a lexicon of 4868 Dutch polysyllabic monomorphematic words (for details see Section 2.1), approximately 80% are regular according to a generally accepted metrical analysis (Trommelen and Zonneveld 1989, 1990). The remaining 20% have to be dealt with in terms of idiosyncratic marking (such as, for instance, exception features or simply a marking of the irregular pattern in the lexicon). On the other hand, the domain exhibits a large number of local ambiguities, or, in other words, it can be said to be noisy. For instance, of the items in the aforementioned lexicon, a metrical encoding (using syllable weights—see below) was performed and it revealed that only 44 of the 89 attested combinations of syllable weights were unambiguous with respect to stress assignment.

In sum, it can readily be seen that the microcosm of metrical phonology is endowed with generalizations as well as irregularities, a phenomenon characteristic of the macrocosm of the linguistic system in general.

1.2 Machine Learning of Metrical Phenomena

Recently, computational learning models that specifically address the problem of learning the regularities of stress assignment have been proposed. These include Gupta and Touretzky (1994), Dresher and Kaye (1990), Dresher (1992), and Nyberg (1991). We will briefly review these models in this section.

Dresher and Kaye (1990) and Nyberg (1991) approach the learning problem from the angle of the Principles and Parameters framework (Chomsky 1981), and they explicitly incorporate the constructs of that theory into their models. It is assumed in this approach that the learner comes to the task of language learning equipped with a priori knowledge incorporated in a universal grammar that constrains him or her to entertain only useful generalizations. More specifically, the a priori knowledge consists of a finite set of parameters, the values of which have to be fixed by the learner. Starting from a finite set of parameters, each with a finite set of possible values, the number of possible grammars that can be developed by the learner is restricted to a finite set.

Computational models such as Dresher and Kaye's (1990) add a learning theory to the (linguistic) notion of universal grammar. This learning theory specifies which aspects of the input data are relevant to each parameter, and it determines how the data processed by the learner are to be used to set the values of the parameters. Eventually, the learner will be able to stress input words, and in doing so will build metrical structures and perform the structure-sensitive operations defined by metrical theory.

Gupta and Touretzky (1994) tackle the problem of learning linguistic stress from a different angle: a simple two-layer perceptron is used as the learning device. In their perceptron model there is no explicit representation of the notion of parameter or the process of parameter setting in any sense. Their system does not aim at setting the correct values of parameters given a learning theory especially designed to do so: "the learning theory employed consists of one of the general learning algorithms common in connectionist modelling." (p. 4) Moreover, their system does not build metrical

representations in the sense proposed in metrical theory when determining the stress pattern of a particular word. Thus, learning in the perceptron is not related in any obvious way to setting the values of parameters that specify the precise geometry of metrical trees. Nor is producing the stress pattern of a particular word related in any obvious way to the construction of a metrical tree and to structure-sensitive metrical operations.

The learning material for Gupta and Touretzky's perceptron consists of the stress patterns of 19 languages.[1] It appears that the learning times for the stress patterns vary according to several dimensions: they describe six dimensions that act as determinants of learnability. For instance, it will take longer for the perceptron to learn the stress pattern of a language that incorporates the factor 'inconsistent primary stress' than to learn a language that does not show that feature. These factors or—so to speak—'parameters' do not coincide with the parameters proposed in metrical theory. However, it is pointed out that there is a close correspondence between ease of learning in the perceptron (as measured by learning times) and some of the markedness and (un)learnability predictions of metrical theory.

The simulations of Gupta and Touretzky show that data-oriented acquisition of stress assignment is possible. Moreover, in observing the perceptron learn stress systems, a number of factors are discovered that appear to determine the learning process. This account of the behavior of the model is termed a 'pseudo-linguistic' theory, and some interesting parallels with metrical phonology are drawn. The crucial point is, however, that the perceptron is not equipped with a priori knowledge about the domain, nor with a specifically designed learning theory.

There are some drawbacks to the simulations presented by both Dresher and Kaye and Gupta and Touretzky. One of the main objections is that they use highly simplified versions of the linguistic data, i.e. small samples encoded using syllable weight only, and without attention to irregularities. Such highly stylized characterizations of stress systems may well capture the core of a language system, but a processing model that aims at learning the stress system of a language should go further. It should also deal with the noise in the actual linguistic data, the irregularities, and the plain exceptions. Gupta and Touretzky (1994:27) appear to be aware of this limitation in their approach:

> "It could be argued that a theoretical account is a descriptive formalism, which serves to organize the phenomena by abstracting away from the exceptions in order to reveal an underlying regularity, and that it is therefore a virtue rather than a failing of the theoretical analysis that it ignores "performance" considerations. However, it becomes difficult to maintain this with respect to a processing model that uses the descriptive formalism as its basis: the processing or learning account still has to deal with actual data and actual performance phenomena."

The research reported in this paper aims at exploring the potential of a learning algorithm that shares the data-oriented (empiricist) mode of learning with the perceptron used in the simulation experiments discussed above, instead of the nativist approach exemplified by the research of Dresher and Kaye (1990). The learning material consists of a lexicon that contains a substantial amount of the attested monomorphemic multisyllabic words of Dutch (see Section 2.1). In this learning material, the details

1 These are the stress patterns of the languages also used by Dresher and Kaye (1990). They represent a selection of the possible stress systems along a variety of metrical dimensions.

of the stress system are not simplified to arrive at a regularized description of the system. Instead, it actually contains the patterns we may expect a language learner to be confronted with.

First, we show that a data-driven alternative to the Principles and Parameters approach is feasible, given a set of examples of a language, in this case Dutch. It is shown that (i) the major generalizations governing main stress assignment can be acquired as well as the major classes of subregularities; and (ii) that the kind of a priori knowledge assumed in the Principles and Parameters approach appears to be unnecessary, even to the extent that the less 'theoretical bias' encoded in the input, the better the learning results are. More specifically, experimental results unequivocally indicate that a phonemic input encoding yields superior results to an encoding in which only the phonological notion of syllable weight is represented.

Secondly, the correspondences of our learning results with metrical theory will be studied: the results of the simulations reveal interesting correlations between learnability by the artificial learner, and markedness in a metrical framework.

Finally, the algorithm's own classification of the test words is analyzed. The algorithm discovers subregularities in the data that are not expressible in metrical terms. Instead it uses the phonemic material presented to form subcategories that act as homogeneous classes with respect to stress assignment. This finding suggests that metrical theory could benefit from proceeding to incorporate segmental information in order to arrive at a more complete description of the data.

The remainder of the paper will be organized as follows: we will first present the most relevant facts about and a metrical analysis of the stress system of Dutch. Next the artificial learning algorithm will be introduced, followed by a discussion of the experimental results.

2. The Problem Domain

2.1 Basic Facts about Dutch Stress Patterns

In this section we will introduce the problem domain, i.e. main stress assignment in Dutch monomorphemic words.[2] In order to do this, we will first discuss the general characteristics of main stress assignment in Dutch and will then proceed to a metrical analysis that adequately captures the generalizations governing the domain.

For the purpose of the experiments to be presented in this paper, we compiled a corpus containing 4,686 polysyllabic monomorphemic words. This corpus was extracted from the CELEX[3] lexical database (Burnage 1990), which contains 130,778 lemmas and 399,186 wordforms and was compiled on the basis of the INL corpus of present-day Dutch (more than 42 million words in a variety of text types). As such, our corpus constitutes a representative sample of Dutch monomorphemes.

In Table 1 and Table 2 these data are divided into bisyllabic and longer words. Within each table, words are divided as to the phonological makeup of their two final syllables, or more precisely their two final rhymes; syllable-initial consonants are not represented. The pattern ə stands for syllables containing a schwa, optionally followed by a consonant word-finally.[4] **VV** denotes a long vowel in an open syllable, while **VC**

2 We are well aware that metrical theory embraces more than main stress assignment in underived words. Yet, as will become clear from this section, this is a far from trivial problem for Dutch. Therefore we limit our attention to this task.

3 Copyright Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

4 For a discussion of the phonemic status of schwa vs. schwa resulting from reduction, see, for example, Trommelen and Zonneveld (1989), Kager (1989), and Zonneveld (1993).

Table 1
Stress patterns in bisyllabic words.

Pattern	PEN			FIN			Row Total
	#	MA	Example IPA Transcription Translation	#	MA	Example IPA Transcription Translation	
ə-ə	0	/	/	0	/	/	0
VV-ə	470	R	tafel ('table') /ta:fəl/	0	/	/	470
VC-ə	494	R	amper ('hardly') /ɑmpər/	0	/	/	494
VXC-ə	35	R	waarde ('value') /wa:rdə/	0	/	/	35
ə-VV	0	/	/	7	R	revue ('review') /rəvy:/	7
VV-VV	201	R	lelie ('lily') /le:li:/	64	LF, [−ex]	cadeau ('present') /kɑdo:/	265
VC-VV	189	R	armoe ('poverty') /ɑrmu:/	40	LF, [−ex]	spondee ('spondee') /spɔnde:/	229
VXC-VV	17	R	extra ('extra') /ɛkstra:/	3	LF, [−ex]	tournee ('tour') /tu:rne:/	20
ə-VC	0	/	/	9	R	rebel ('rebel') /rəbɛl/	9
VV-VC	177	R	epos ('epic') /e:pɔs/	135	[−ex]	bizar ('bizarre') /bi:zɑr/	312
VC-VC	135	R	cactus ('cactus') /kɑktʉs/	101	[−ex]	trompet ('trumpet') /trɔmpɛt/	236
VXC-VC	8	R	oorlog ('war') /o:rlɔχ/	5	[−ex]	transfer ('transfer') /trɑnsfɛr/	13
ə-VXC	0	/	/	21	R	reflex ('reflex') /rəflɛks/	21
VV-VXC	45	I	climax ('climax') /kli:mɑks/	384	R	alarm ('alarm') /a:lɑrm/	429
VC-VXC	43	I	potlood ('pencil') /pɔtlo:t/	334	R	albast ('alabaster') /ɑlbɑst/	377
VXC-VXC	6	I	argwaan ('suspicion') /ɑrχwa:n/	20	R	punctuur ('puncture') /pʉnkty:r/	26

stands for a closed syllable containing a short vowel followed by a single consonant. Since Dutch lacks short vowels in open syllables, a single intervocalic consonant is taken to be ambisyllabic when following a short vowel, thus yielding a VC-pattern (cf. van der Hulst 1984). The pattern **VXC** abbreviates both a long vowel followed by at least one consonant and a short vowel followed by at least two consonants. This type of syllable is usually restricted to final position. The column labels **ANT**, **PEN**, and **FIN** denote *antepenultimate*, *penultimate*, and *final* stress respectively. These columns provide information about the frequency of each type of stress for each phonological pattern. The columns headed by the label MA will be discussed later; they need not concern us at this point.

From these tables it appears that the three possible stress patterns occur with different frequency: PEN is the most frequent pattern (52.96% of all words), ANT is the least frequent pattern (7.46% of all words), and FIN is in between (39.58% of all

Table 2
Stress patterns of trisyllabic and longer words.

Pattern	ANT			PEN			FIN			Row Total
	#	MA		#	MA		#	MA		
-ə-ə	1	R	gisteren ('yesterday') /ɣɪstərə/	0	/	/	0	/	/	1
-VV-ə	9	R	maluwe ('malve') /ma:ly:wə/	219	R	bagage ('luggage') /ba:ɣa:ʒə/	0	/	/	228
-VC-ə	0	/	/	91	R	amandel ('almond') /a:mɑndəl/	0	/	/	91
-VXC-ə	0	/	/	8	R	/	0	/	/	8
-ə-VV	10	LF	opera ('opera') /o:pəra:/	0	/	/	21	LF, [−ex]	selderie ('celery') /sɛldəri:/	31
-VV-VV	125	LF	eskimo ('Eskimo') /ɛski:mo:/	212	R	calvarie ('calvary') /kalva:ri:/	91	LF, [−ex]	fantasie ('fantasy') /fɑnta:zi:/	428
-VC-VV	2	I	penalty ('penalty') /pɛnɑlti:/	126	R	placenta ('placenta') /pla:sɛnta:/	34	LF, [−ex]	anarchie ('anarchy') /a:nɑrχi:/	162
-VXC-VV	0	/	/	1	R	balalaika ('balalaika') /ba:lɑ:la:jka:/	2	LF, [−ex]	chevalier ('horseman') /ʃəva:lje:/	3
-ə-VC	19	R	boemerang ('boomerang') /bu:mərɑŋ/	0	/	/	19	[−ex]	borderel ('statement') /bɔrdərɛl/	38
-VV-VC	170	R	bariton ('baritone') /ba:ri:tɔn/	64	LF	dictator ('dictator') /dɪkta:tɔr/	67	[−ex]	minaret ('minaret') /mi:na:rɛt/	301
-VC-VC	3	I	badminton ('badminton') /bɑtmɪntɔn/	36	R	universum ('universe') /y:ni:vɛrsʉm/	12	[−ex]	bombardon ('bombardon') /bɔmbɑrdɔn/	51
-VXC-VC	0	/	/	0	/	/	0	/	/	0
-ə-VXC	12	I	kandelaar ('candlestick') /kɑndəla:r/	0	/	/	78	R	arsenaal ('arsenal') /ɑrsəna:l/	90
-VV-VXC	11	I	olifant ('elephant') /o:li:fɑnt/	0	/	/	404	R	magistraat ('magistrate') /ma:gi:stra:t	416
-VC-VXC	1	I	leukoplast ('leucoplast') /lɸ:ko:plɑst/	1	I	appendix ('appendix') /apɛndɪks/	75	R	resultaat ('result') /re:zʉlta:t/	77
-VXC-VXC	0	/	/	0	/	/	1	R	conjunctuur ('conjuncture') /kɔnjʉnkty:r/	1

patterns). A first glance at both tables might suggest almost arbitrary variation of main stress. But a number of near exceptionless generalizations can be formulated (see also Kager 1989):

(i) Main stress is restricted to a three-syllable window from the right-hand word edge.

(ii) Syllables containing a schwa are never stressed; moreover, stress almost always falls on the immediately preceding syllable.

(iii) Antepenultimate main stress may occur if the penult is a VV-syllable, but apart from only a few exceptions, never with a VC-syllable.

Apart from these observations, there are a number of general tendencies worth mentioning.

(i) Final VXC syllables tend to attract main stress, in both bisyllabic and longer words.

(ii) In other bisyllabic words, penultimate stress is the dominant pattern, although final stress is more common in VX-VC words than in words ending in an open syllable.

(iii) In trisyllabic and longer words, VC-final words tend to have stress on the antepenultimate syllable, if the penult is open, and stress on the penult if it is closed. For VV-final words, penultimate stress is the dominant pattern, regardless of the structure of the penult; final stress in these words does occur, but is more uncommon than antepenultimate stress.

Given this description of the data, the challenge for a theoretical analysis is both to capture the relevant generalizations in a natural way and to provide a principled account for the relative markedness of nondominant patterns. We will turn to an analysis that meets both requirements.

2.2 A Metrical Analysis of Dutch Stress Assignment

The theoretical analysis of main stress assignment we will present in this section is cast in the framework of metrical phonology, a branch of nonlinear phonology that is concerned with phonological constituency and the prominence relations that hold between categories at various hierarchical levels.[5] Dutch stress has been the subject of a lively discussion during the last decade (for an overview see Trommelen and Zonneveld [1989] and Kager [1989]). We will briefly sketch the analysis of Trommelen and Zonneveld. It is not only the most fully articulated tree-based analysis of Dutch stress to date, but also represents what has since become the consensus view.[6]

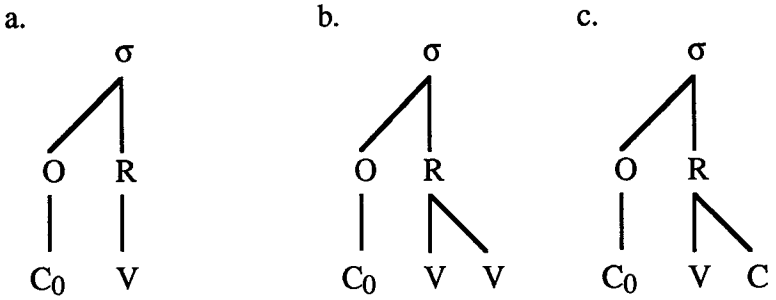
The focus will be on how the regularities of the stress system are captured and how markedness is related to nondominant patterns. We will pay special attention to those places where Dutch deviates from the universal 'default.'

5 For an introduction to metrical phonology, see, e.g., Goldsmith (1990) and van der Hulst and Smith (1982).

6 Although in more recent work by Kager (1989) and Zonneveld (1993) the analysis is restated in a formalism using bracketed grids (see Halle and Vergnaud [1987]), the main insights from Trommelen and Zonneveld (1989) concerning the nature and amount of lexical markings needed are retained.

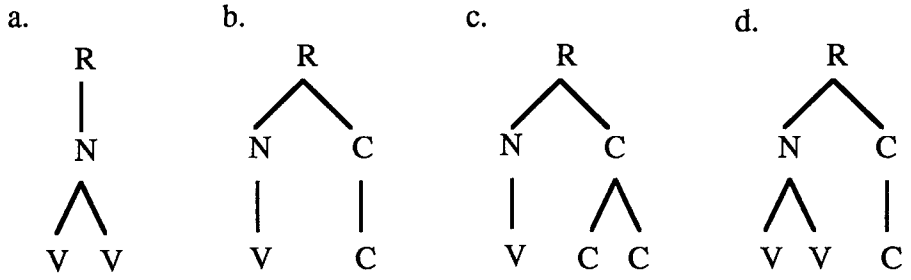
2.2.1 Syllable Structure and Foot-Building. In Tables 1 and 2 four different rhyme templates were distinguished, ə, VV, VC and VXC, and it was shown that each one exhibited different stress properties. In various analyses of Dutch these types have been referred to as superlight (ə), light (VV), heavy (VC), and superheavy (VXC). It is not uncommon that stress systems are sensitive to syllable structure, and languages that distinguish between light and heavy syllables are known as quantity-sensitive. Yet the particular distinction proposed for Dutch merits some further discussion since cross-linguistically, a VV rhyme counts as heavy whenever a VC rhyme does—an observation that does not seem to hold for Dutch. In a tree-based framework, syllable weight has been related to the degree of branching of the rhyme, in the following manner (see Example 1). In what follows, σ will be used as the label of a syllable node, and O and R are abbreviations for Onset and Rhyme, respectively. N stands for Nucleus and when opposed to N, C is used to denote Coda.

Example 1



The dividing line between light and heavy syllables is usually drawn between (a) and (b). However, in her study of Dutch syllable structure, Trommelen (1983) argues that for Dutch a further distinction is needed, in which the rhyme is analyzed as consisting of a peak (also called nucleus), containing the vocalic part, and a coda, containing any remaining consonants. Further, she argues that Dutch rhymes can exhibit the structures shown in Example 2, where (c) and (d) are restricted to the word edge:

Example 2



Since Dutch lacks short vowels in open syllables, mere branching of the rhyme is insufficient to establish a weight distinction between syllables. Trommelen (1983) therefore proposes that in Dutch this notion has to be replaced by that of direct vs. indirect branching of the rhyme. Thus a rhyme that branches directly into a peak and a coda counts as heavy, whereas a branching peak only does not. In other words, the weight distinction in Dutch between *light* and *heavy* syllables seems to coincide with the distinction between *open* and *closed* syllables.

In a further elaboration of Trommelen's analysis, Kager and Zonneveld (1986) focus on 'superheavy' and 'superlight' syllables. The excess consonant(s) in superheavy syllables are analyzed as an extrasyllabic appendix, which is restricted to domain edges. More importantly, word-final schwa syllables are given the same treatment, based on arguments distinct from their stress properties. The net effect of this analysis is that preceding consonants are pushed onto the previous rhyme, making this rhyme (super)heavy.

The relevance of these observations for stress assignment lies in their impact on foot formation. The rules for foot formation are the following:

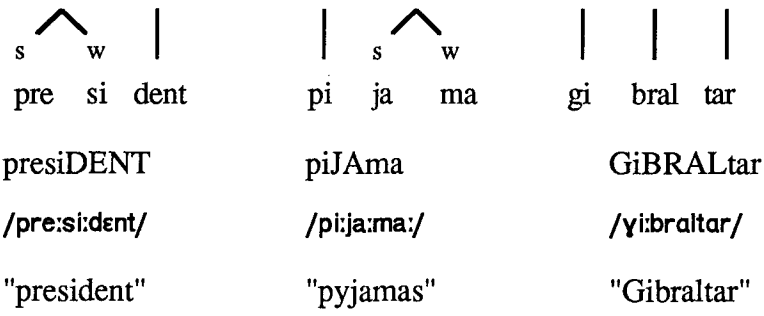
- Construct maximally binary feet, going from right to left.
- Feet are labeled s-w.

Quantity-sensitivity (Q.S.) in the sense defined above allows the reformulation of a universal restriction for Q.S. languages, i.e. that heavy syllables cannot occur in weak foot position (Hayes 1981), in the following manner:

- Closed syllables may not occur in recessive (weak) foot position.[7]

This restriction leads to the creation of monosyllabic feet over VC and VXC syllables. Some relevant examples are shown in Example 3.

Example 3



2.2.2 Word Tree Labeling. Whereas the foot-building conventions above have been relatively uncontroversial since Kager (1985), word-tree formation has raised considerably more controversy. We will now discuss Trommelen and Zonneveld's (1990) proposal, the approach that we will adhere to in what follows.

7 This formulation was first made by Kager (1985).

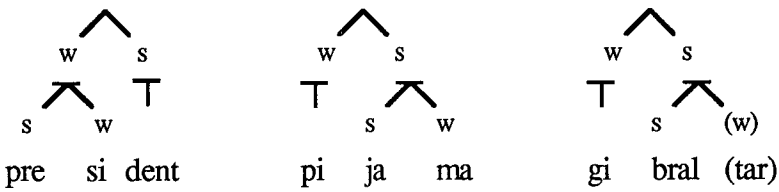
Their starting point is to build a uniformly right-branching right-dominant word tree that locates main stress on the final foot.[8] This makes the right predictions for VV and VXC final words, i.e. words having stress on the penultimate and final syllable respectively. Obviously, to generate stress on other positions, additional mechanisms are needed. Antepenultimate stress, which is the dominant pattern in -VV-VC final words, cannot be achieved if stress consistently falls within the final foot. Conversely, to account for final stress in VV-final words, a monosyllabic foot seems required to provide a landing site for the end rule.

The devices that seem called for are the following: on the one hand, the possibility of assigning a lexically prespecified monosyllabic foot (henceforth abbreviated as LF) to cover the VV-final words with final stress, and on the other hand, extrametricality to handle antepenultimate stress. Extrametricality amounts to making an element invisible to stress rules and is restricted to domain edges under the Peripherality Condition (Hayes 1981). Since in VC-final words final stress is not the dominant pattern, assignment of [+ex] for this type of word should be rule-governed, rather than an exceptional marking. The rules for word tree formation then are the following:

- Mark a final VC-rhyme as extrametrical before foot formation applies.
- Construct a right-branching word tree, labeled uniformly w-s.

This leads to the following trees for the regular patterns, where the word tree is drawn above the horizontal marks. Note that in 'gibraltar' the final syllable is marked as extrametrical before foot formation applies, and only later incorporated as a weak foot member by a universal convention of Stray Syllable Adjunction (Hayes 1981).

Example 4

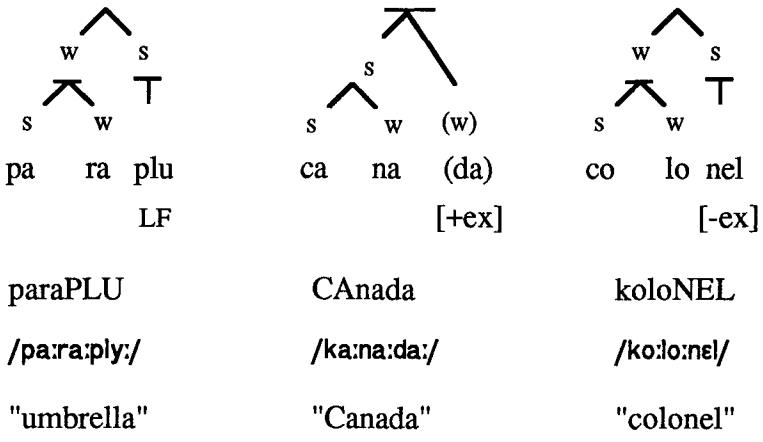


presiDENT	piJAma	GiBRALtar
/pre:si:dɛnt/	/pi:ja:ma:/	/ɣi:braltar/
"president"	"pyjamas"	"Gibraltar"

8 This rule is therefore often referred to as the End Rule.

Nondominant patterns are handled with three exception features: a lexical foot (LF) for VV-finals with final stress, [+ex] for VV-finals with antepenultimate stress, and [−ex] for VC-finals with final stress. The trees in Example 5 illustrate the analysis for these marked patterns:

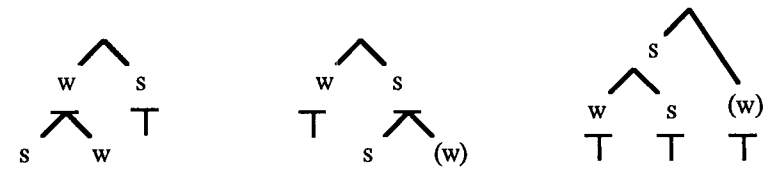
Example 5



This situation is clearly not ideal, for several reasons: first, it is hard to account for degrees of markedness when three exception mechanisms are at play. For example, final stress is more common in VX-VC words than it is in VX-VV words, yet relating this to [−ex] vs. LF respectively does not bring out this contrast. Another objection comes from stress shifts and mispronunciations. Van Marle (1978) adduces various kinds of evidence, of which the most illuminating examples are words like 'rococo.' Rococo is attested both with final stress (as in 'paraPLU') and with antepenultimate stress (as in 'CAnada'). Under the current account, it is hard to explain how the loss of the LF feature in the case with final stress would automatically imply adoption of the [+ex] feature, to yield antepenultimate stress, rather than producing the unmarked penultimate pattern. A third problem is that by making a final VC syllable consistently extrametrical, the important generalization about its footing behavior both word-finally and word-internally is lost.

The solution Trommelen and Zonneveld (1990) propose for these problems is that, in principle, nothing prevents extrametricality from applying 'late,' i.e. after foot formation. Extrametricality would then affect the word tree only, and leave the foot formation rules untouched. A further modification they propose is to mark all final VX syllables as extrametrical, including final VV. The cases handled correctly by the end rule are still regular, since final VXC syllables are exempt from extrametricality. Also, because extrametricality is not allowed to percolate from a nonhead position, for VV-final words the default is still penultimate stress. The trees in Example 4 are modified now, to yield the trees in Example 6 for the regular cases:

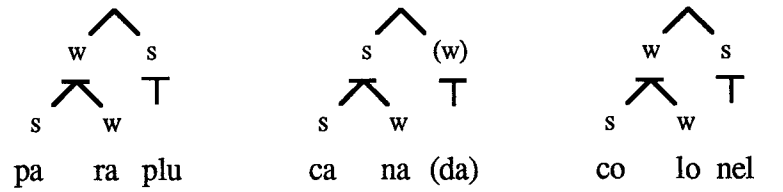
Example 6



presiDENT	piJAma	GiBRALtar
/pre:si:dɛnt/	/pi:ja:ma:/	/ɣi:brɑltɑr/
"president"	"pyjamas"	"Gibraltar"

For the marked patterns, lexical feet are retained, but exceptional extrametricality is no longer needed, since antepenultimate stress now results from such an idiosyncratically specified foot. The other problems are solved, since the opposition between VX-VC and VX-VV words amounts to the difference between LF and [−ex] vs. [−ex] by itself. Stress shifts of the type 'rocoCO'–'ROcoco' are explained by loss of the feature [−ex] only, where LF is retained (compare the trees for 'paraPLU' and 'CAnada' in Example 7 below). Hence, in their final analysis the patterns in Example 5 are analyzed as in Example 7.

Example 7



paraPLU	CAnada	koloNEL
/pa:ra:ply:/	/ka:na:da:/	/ko:lo:nɛl/
"umbrella"	"Canada"	"colonel"

We can now return to Table 1 and Table 2. In the columns headed MA ('Metrical Analysis'), the lexical markings of the various patterns are indicated. In these tables,

R stands for regular, **LF** for a lexically prespecified foot, **[−ex]** for an exception to rule-governed extrametricality, and **I** for unexplained exceptions that need full lexical marking. It appears that out of the 4868 words in our lexicon, 81.1% are regular (R), 6.96% are exceptions to the extrametricality condition ([−ex]), 4.07% require a specified lexical foot (LF) and 5.24% a combination of the two preceding features (LF, [−ex]), and finally 2.59% are plain exceptions (I). These five categories can be scaled according to their markedness within the metrical framework: the regular case (R) is of course the least marked, the irregular (I) the most marked. In between these extremes, one single exception feature is less marked than two exception features, i.e., items that need an LF and a [−ex] are more marked than items that are either marked as LF or [−ex].

In this section we have sketched a metrical analysis of the Dutch stress system that captures the relevant generalizations in a natural way and provides a principled account for the relative markedness of nondominant patterns. The dominant patterns are rule-generated, while deviations from this pattern are handled by two types of lexical marking. Cumulation of these markings accounts for degrees of exceptionality and explains why stress shifts do not always move in the direction of the dominant pattern.

In the analysis, it was pointed out that Dutch is fairly idiosyncratic in a number of ways: first, the weight distinction between VV and VC is odd from a universal perspective, and secondly, extrametricality in Dutch influences the word tree only. Furthermore, Dutch makes liberal use of lexical markings, and this has led Kager (1989) to conclude that Dutch, while not being a free stress language, occupies a middle ground between free and fixed stress systems.

3. The Learning Algorithm

Assigning stress to a word can be interpreted as a classification problem: given a pattern (a set of feature-value pairs describing a word), the task of the system is to decide whether stress is on the final (FIN), penultimate (PEN), or antepenultimate (ANT) syllable. In other words, the system has to decide whether the word belongs to category FIN, PEN, or ANT. Notice that we are only trying to predict *main stress*; for predicting secondary stress or different stress levels, a more elaborate category system has to be used.

3.1 Instance-Based Learning

The data-oriented algorithm we used is a variant of Instance-Based Learning (IBL, Aha, Kibler, and Albert 1991). IBL is a framework and methodology for incremental, supervised, similarity-based learning.

> **Supervised**. The system is trained by presenting a number of patterns with their classification.
>
> **Incremental**. Training material can be added one item after the other, without a need to retrain the system (unlike, e.g., backpropagation in connectionist networks, where batch learning is used).
>
> **Similarity-Based**. The system bases its category prediction on the similarity of an unseen test pattern to the training patterns to which it was exposed earlier. The distance metric used to compare the similarity of patterns is therefore the most important aspect of the algorithm.

In research described elsewhere (Gillis et al. 1992; Daelemans et al. 1993), we experimented with two additional learning algorithms: Backpropagation of Errors in Feedforward Networks (Rumelhart, Hinton, and Williams 1986), and Analogical Modeling (Skousen 1989). Although there are small differences in the learning behavior of systems trained with these different learning algorithms on the task of stress assignment, the overall performance of the systems was highly similar. We therefore decided to limit our attention to IBL, which is the simplest and most transparent of the three learning algorithms.

The distinguishing feature of IBL is the fact that no explicit abstractions are constructed on the basis of the training examples during the training phase. A selection of the training items themselves is used to classify new inputs. IBL shares with Memory-Based Reasoning (Stanfill and Waltz 1996) and Case-Based Reasoning (Riesbeck and Schank 1989) the hypothesis that much of intelligent behavior is based on the immediate use of stored episodes of earlier experience rather than on the use of explicitly constructed abstractions extracted from this experience (e.g. in the form of rules or decision trees). In the present context of learning linguistic tasks, the hypothesis would be that much of language behavior is based on this type of memory-based processing rather than on rule-based processing. In linguistics, a similar emphasis on analogy to stored examples instead of explicit but inaccessible rules, is present in the work of, among others, Derwing and Skousen (1989).

IBL is inspired to some extent by psychological research on exemplar-based categorization (as opposed to classical and probabilistic categorization, Smith and Medin [1981], Nosofsky, Clark, and Shin [1989]). Finally, as far as algorithms are concerned, IBL finds its inspiration in statistical pattern recognition, especially the rich research tradition on the nearest-neighbor decision rule (see, e.g., Devijver and Kittler [1982] for an overview).

3.2 The Algorithm

It is useful to distinguish between a *learning component* and a *performance component* when describing learning systems. The performance component carries out the required task (in this case predicting the stress category of unseen words), and the learning component changes the system in response to the examples presented (in the case of IBL by simply storing the examples) such that the accuracy of the system increases.

3.2.1 Training. During training, pre-categorized training items are presented in an incremental fashion to the learning component. A training item is a sequence of feature-value pairs (for instance, a sequence of the weights associated with a word's syllables) with its category (in this case, the stress category of a word). If the pattern was not encountered earlier, a new memory record is created, listing the pattern and initializing its *category distribution* (a record showing for each possible category the number of times the pattern was associated with this category in the training set). As the same pattern may represent different words, depending on the encoding used, the category distribution contains probabilistic information about the category of ambiguous patterns (patterns that are assigned different categories in the training set). If the training pattern was encountered earlier during training, its category distribution is updated.

3.2.2 Testing. The operation of the performance component of the IBL algorithm is quite simple: for each test item (a sequence of feature-value pairs to be assigned a category), we check whether it is present in memory. If this is the case, the category assigned most often to this pattern (as evidenced in its category distribution) is as-

signed to the test item. If the test item has not yet been encountered, its similarity to all patterns kept in memory is computed, and a category is assigned based on the category of the most similar item(s).

Similarity is measured using a distance metric: two patterns are similar if their distance in pattern space is small. If there is only one best match, the most frequent category in its category distribution is used. If there is a tie between two or more patterns in memory, their category distributions are combined (summed) before the most frequent category is selected as the category predicted for the test item.

The performance of an IBL classifier crucially depends on the selection of training items to be kept in memory and the distance metric used. In the experiments described here, we 'remembered' all training items. The distance metric will be elaborated on in what follows.

The most straightforward distance metric would be the one in equation (1), where X and Y are the patterns to be compared, and $\delta(x_i, y_i)$ is the distance between the values of the i-th feature in a pattern with n features.

$$\Delta(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{1}$$

Distance between two values is measured using equation (2) for numeric features and equation (3) for symbolic features.

$$\delta(x_i, y_i) = \frac{|x_i - y_i|}{\max_i - \min_i} \tag{2}$$

$$\delta(x_i, y_i) = 0 \text{ if } x_i = y_i, \text{ else } 1 \tag{3}$$

When computing the distance between numeric values, dividing by the difference between the maximum and minimum values of a feature scales numeric features with different lower and upper bounds to comparable differences between 0 and 1.

3.2.3 Information Gain. When using a geometrical distance metric for numeric features (geometrical distance between two patterns in pattern space), or an overlap metric for symbolic features (number of features with equal values in both patterns), all features are interpreted as being equally important. But this is of course not necessarily the case. We extended the basic IBL algorithm proposed by Aha, Kibler, and Albert (1991) with a technique for automatically assigning a different importance to different features. Our approach to the problem of weighing the relative importance of features is based on the concept of Information Gain (IG, also used in learning inductive decision trees [Quinlan 1986]), and first introduced (as far as we know) in IBL in Daelemans and van den Bosch (1992) in the context of a syllable segmentation task. The idea is to interpret the training set as an information source capable of generating a number of messages (the different categories) with a certain probability. The information entropy of such an information source can be compared in turn for each feature with the average information entropy of the information source when the value of that feature is known.

Database information entropy is equal to the number of bits of information needed to know the category given a pattern. It is computed by the formula in (4), where p_i (probability of category i) is estimated by its relative frequency in the training set.

$$H(D) = -\sum_{i} p_i \log_2 p_i \tag{4}$$

For each feature (position in the patterns), we now compute what the information gain is in knowing its value. To do this we have to compute the average information entropy for this feature and subtract it from the information entropy of the database. To compute the average information entropy for a feature, we take the average information entropy of the database restricted to each possible value for the feature. The expression $D_{[f=v]}$ refers to those patterns in the database that have value v for feature f; V is the set of possible values for feature f.

$$H(D_{[f]}) = \sum_{v_i \in V} H(D_{[f=v_i]}) \frac{|D_{[f=v_i]}|}{|D|} \tag{5}$$

Information gain is then obtained by equation (6) and scaled to be used as a weight for the feature during similarity matching.

$$G(f) = H(D) - H(D_{[f]}) \tag{6}$$

The distance metric in equation (1) is then modified to take into account the information gain weight associated with each feature.

$$\Delta(X, Y) = \sum_{i=1}^{n} G_i \delta(x_i, y_i) \tag{7}$$

To retain the incremental character of IBL, we updated the information gain weights with every new training item. For the present task, the weights hardly change after about 100 training patterns, and further changes have no effect on performance.

4. Experiment

Having introduced the problem domain and the learning algorithm, we are ready to discuss the results of the experiment on stress assignment. For this task, words (training and test patterns) were represented by three different feature-value pair encodings, which will be discussed in the next section. Output of the system consists of a prediction of the category (FIN, ANT, or PEN) of the input word. Actually, more detailed information is provided: by using the category distribution described earlier, for each possible category, a value between 0 and 1 representing the probability that the word has this category can be provided. However, no use was made of this in the experiment. A single output category is selected for each pattern: the one with the highest probability or a random choice in case of a tie.

The main aims of the experiment are (i) to assess the role of the encoding used, and more specifically, to investigate the impact of 'theoretical bias' in the input encodings on the learning success, and (ii) to relate the learning performance of the algorithm to the metrical analysis of the previous section.

4.1 Method

The method used in this experiment consisted of a ten-fold cross-validation experiment (Weiss and Kulikowski 1991). In this set-up, the database is partitioned ten times, each with a different 10% of the dataset as the test part. The remaining 90% is used as the training part.

For each of the ten simulations in our experiment, the test part was used to test generalization performance. The size of the training set was varied from 500 to 4000 items randomly chosen from the training part in order to assess the system's learning performance.

4.2 Data and Data Encoding

In the experiment we used the lexicon of 4868 Dutch multisyllabic monomorphemes introduced in Section 2. In order to use test sets of equal magnitude in the ten-fold cross validation experiments, 8 items were randomly selected from the lexicon and withdrawn from the experiment, so that 10 test sets of 486 items were constructed.

In order to investigate the impact of the input encodings, three encoding schemes were implemented. In each instance only the three last syllables are encoded.

> **Encoding 1**. Strings of syllable weights of the last three syllables of a word, i.e., the kind of encoding judged to be necessary and sufficient for learning a quantity-sensitive language (Dresher and Kaye 1990; Gupta and Touretzky 1993);
>
> **Encoding 2**. The phonemic information contained in the rhyme projections of the last three syllables;
>
> **Encoding 3**. A plain phonemic transcription of the word.

Encoding 1 is based on the notion of syllable weight and uses a single feature for each syllable. Since in metrical phonology syllable weight is a function of the degree of branching of the rhyme, the set of values chosen should discriminate among different rhyme types. We used numerical values, ranging from one to five, to set up a weight scale (see Section 2.2.1 for a discussion of the weight scale) in the following manner:[9]

1 = superlight rhymes (rhymes containing a schwa, ə);
2 = light rhymes (a long vowel in an open syllable, VV);
3 = heavy rhymes (a short vowel followed by a single consonant, VC);
4 = superheavy rhyme of the type VCC;
5 = superheavy rhyme of the type VVC.

The word *agenda* ('agenda,' IPA transcription: /a:ɣɛnda:/) is encoded as the sequence '2 3 2', i.e., a light syllable (2) followed by a heavy one (3) and a light one (2). Thus in this encoding only three features are used. The value of the first feature is the syllable weight of the antepenultimate syllable, the value of the second feature, the weight of the penultimate syllable, and the value of the third the weight of the final syllable.

Encoding 2 provides a phonemic encoding of the rhyme and uses two features per syllable, one for the nucleus and one for the coda. It coincides with the previous encoding in the sense that it too provides the necessary information on which syllable weight is based, albeit without abstracting over phonemic detail as was done in Encoding 1. Thus the encoding for the word *agenda* ('agenda') looks as follows: a: - ɛn a: -.[10] The first syllable has nucleus /a:/ and an empty coda, denoted with a dash; the second syllable nucleus /ɛ/ and coda /n/. The last syllable has nucleus /a:/ and an empty coda. Thus, in the second encoding, six features are used. The first two features stand for the nucleus and the coda of the antepenultimate syllable, the next two features stand for the nucleus and the coda of the penultimate syllable, and the last two features stand for the nucleus and the coda of the final syllable. The values of the features are the phonemes or phoneme strings that occupy these respective positions in the word.

9 See Visch and Kager (1984) for a discussion of why VVC should be "heavier" than VCC.
10 The actual encodings are in DISC format (see Burnage [1990]), which has the advantage that each phoneme is transcribed by means of a single symbol.

Table 3
Sample encodings using the three encoding schemes.

Encoding Number	Target	Word	Encoding
1	PEN	agenda	2 3 2
2	PEN	agenda	a: - ɛn a: -
3	PEN	agenda	- a: - ɣɛn d a: -

Encoding 3 extends the rhyme encoding by adding a feature for the onset of each syllable. As such, it consists of a complete phonemic encoding of the last three syllables. Thus, each of the last three syllables of a word is represented by three features, the values of which represent the phonemes that fill the onset, nucleus, and coda slot of the syllable. For instance, the encoding for the word *agenda* ('agenda') looks as follows: - a: - ɣɛn d a: - . The onset of the first (or antepenultimate) syllable is empty, hence a dash in the encoding, the nucleus is /a:/, and the coda is empty. The second syllable consists of onset /ɣ/, nucleus /ɛ/, and coda /n/. The last syllable consists of onset /d/, nucleus /a:/, and an empty coda.

The three encodings for the word *agenda* ('agenda) are given in Table 3.

It should be noted that the more detailed the representation is (Encoding 3 being the most detailed), the less ambiguous the training patterns are. On the other hand, the more detailed the encoding, the more irrelevant features are presented to the learning system (the data are noisier from the point of view of linguistic theory). Onsets are of little or no use for stress assignment according to metrical analyses, but they are present in the third encoding, thus adding (allegedly) irrelevant information.

The success rate of the algorithm is obtained by calculating the average accuracy (number of test pattern categories correctly predicted) over the ten test sets in the ten-fold cross validation experiment.

4.3 Results

4.3.1 Analysis of General Performance. In this section we will discuss the performance of the algorithm at a general level. The most striking result is that IBL, when trained with phonemic encodings (Encoding 2 and Encoding 3), yields significantly better results than when trained with Encoding 1, the weight string representation.

In Figure 1 results for the three encodings are plotted. Overall peak success scores for the three encodings lie between 80% and 90%, ranging from 81.26% for Encoding 1 and 88.11% for Encoding 2 to 88.81% for Encoding 3. Considering the fact that in a theoretical analysis about 80% of the data was considered regular and that perfect predictions are beyond reach for the Dutch stress system, these figures indicate that the algorithm has picked up the regularities governing the field, and this for all three encodings. This does not mean, however, that the three encodings are equally good.

First of all, the results for Encoding 1 are significantly lower than those for the two other encodings. An analysis of variance (ANOVA) performed over the pooled results per encoding shows a highly significant difference between the results for the three encodings ($F(2, 237) = 286.0978, p < .0001$). Paired t-tests indicate that there is a highly significant difference between the results for Encoding 1 and the two other encodings (Encoding 1–Encoding 2: $t = 20.1985$, $df = 158$, $p < .0001$; Encoding 1–Encoding 3: $t = 22.7998$, $df = 158$, $p < .0001$). In the same vein the difference between the results for the second and the third encodings were calculated, showing a significant

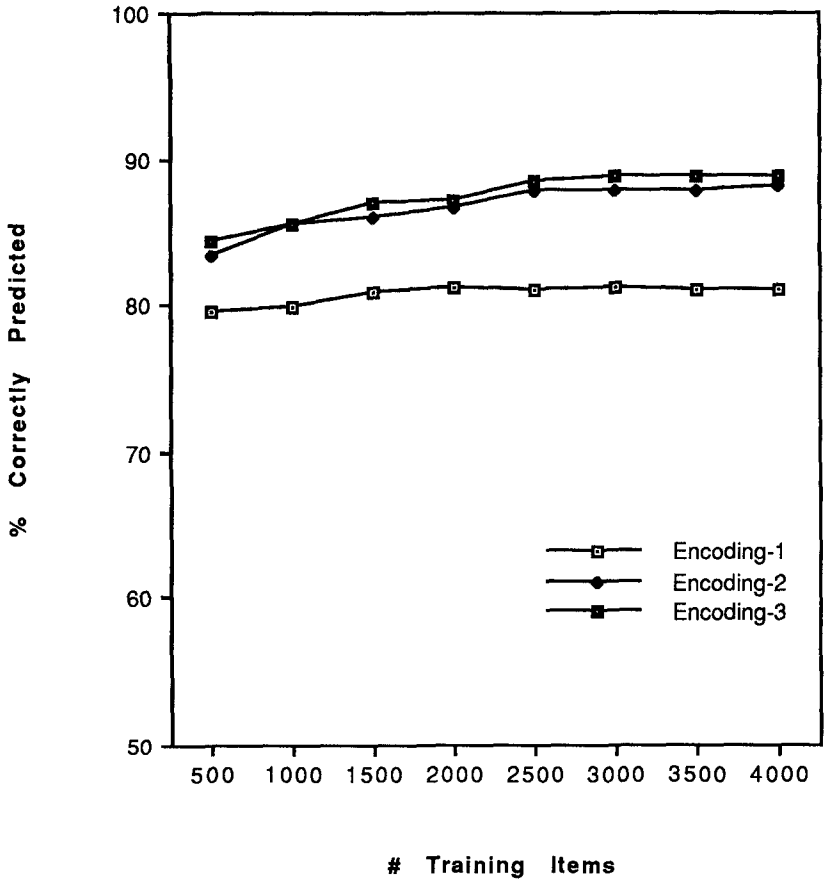


Figure 1
General comparison of success rates.

difference between the two, i.e., Encoding 3 yields significantly higher success rates than Encoding 2 ($t = 2.2792$, $df = 158$, $p < .02$).

A comparison of the peak success scores in the three conditions reveals that Encoding 1 scores significantly less well than the other two (Encoding 1–Encoding 2: $\chi^2 = 149.803$, $p < .01$; Encoding 1–Encoding 3: $\chi^2 = 181.955$, $p < .01$), while the peak score for Encoding 3 is not significantly better than the peak score for Encoding 2 ($\chi^2 = 2.27$, $p > .05$). This shows that the weight string encoding leads to significantly poorer results than the two other encodings.

A second piece of evidence comes from an analysis of the performance of the classifier with regard to the specific target categories. In Table 4 the peak success scores for the individual target categories are displayed.[11]

11 In this table the highest success score per target category is displayed regardless of the number of training items involved to reach this peak score. For instance, the highest score for the target category 'penultimate stress' is reached with 3000 items in the Encoding 1 condition, 3500 items in the Encoding 3 condition, and 4000 items in the Encoding 2 condition. Since we are not mainly concerned with an analysis of the learning curves in these various conditions, these differences will not be of any further concern.

Table 4
Highest success rates for the three encodings relative to target categories.

Encoding Number	FIN	PEN	ANT
1	74.90	93.60	53.19
2	87.94	92.20	61.77
3	89.00	92.93	62.05

These results show that stress on the penultimate syllable, which is the case for 52.96% in the corpus used for training, is learned best with peak success rates varying from 92.20% for Encoding 2 to 93.60% for Encoding 1. Stress on the antepenultimate syllable, which is found in only 7.46% of the lexicon used for training, seems much harder to predict, with peak success rates ranging from 53.19% for Encoding 1 to 62.05% for Encoding 3. Stress on the final syllable (39.58% of the corpus) is predicted correctly in 74.9% of the cases for Encoding 1, in 87.94% for Encoding 2, and in 89.0% for Encoding 3.

Thus, Encoding 1 was shown to lead to lower success rates in the global comparison of the experimental results. The data per target category provide the following picture. It is still the case that the peak performances of Encoding 2 and Encoding 3 do not differ significantly for the three target categories. Moreover, they are both significantly better than Encoding 1 for stress on the final and the antepenultimate syllable. However, Encoding 1 yields equally high (even better) results than the two other encodings when the penultimate syllable is the target category.

In order to fully appreciate the results for penultimate stress, we tested whether IBL shows a tendency to select the most frequent class as an appropriate response, and to overgeneralize that response, a phenomenon not uncommon for statistical learning algorithms. In Figure 2 we plot out the number of times penultimate stress (PEN), the most frequent stress pattern, is predicted by IBL (the total test set consists of 2575 words with PEN as their category). The graph clearly shows that the three encodings have a tendency to generate more PEN responses than there are actual PEN targets in the test set from the very start (training set of 500 items). However, it also clearly shows that Encoding 1 does not behave as Encoding 2 and Encoding 3: in contrast with the two other encodings, the number of PEN predictions increases to arrive at more than 3000 PEN predictions from a training set size of 2000 items onward. This means that the algorithm has found a generalization, and overgeneralizes it because of the low discriminatory ability of the encoding used: at the same time that Encoding 1 shows an increase of PEN answers, Encoding 2 and Encoding 3 seem to offer enough information to make more fine-grained distinctions, so that the overgeneralization of the PEN response is minimized.

Another way of showing this same effect is by analyzing the confusion matrices. We selected the results of the ten-fold cross-validation experiments with a training set of 4000 items, i.e., the largest training set. Confusion matrices were drawn from these results (see Table 5). These matrices should be read as follows: the vertical dimension gives the target category and the horizontal dimension the predicted category. So, the first matrix (with the data from IBL trained with Encoding 1) shows in its upper row the classification of the words that have final stress (FIN). It appears that the classifier predicted this outcome correctly in 70.37% of the cases (upper left cell). However, IBL also predicted the outcome penultimate stress (PEN) in 25.26% of the words and

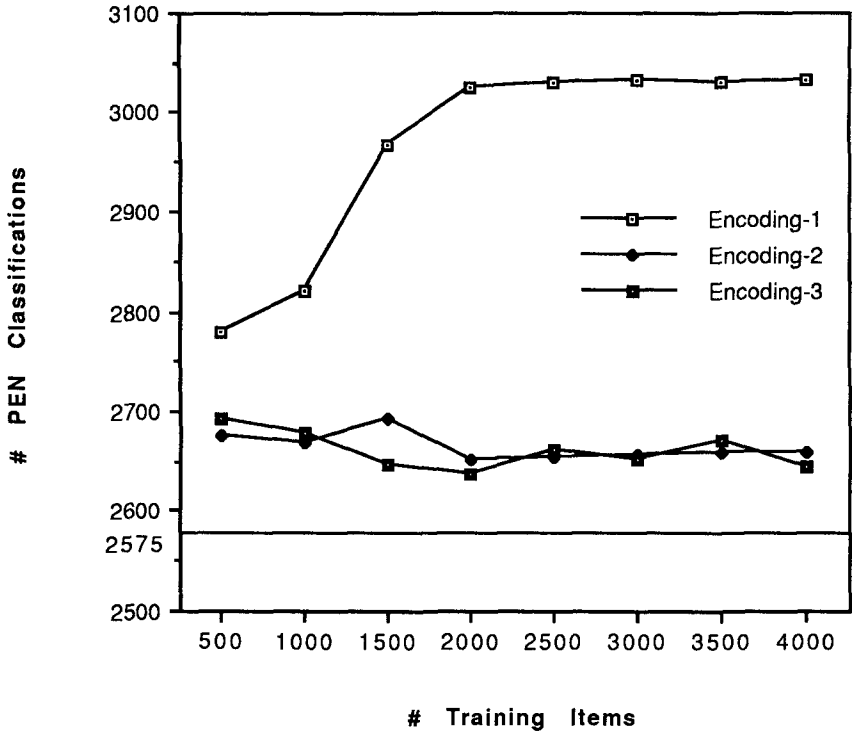


Figure 2
Number of PEN classifications.

Table 5
Confusion matrices for IBL trained with 4000 items (row percentages between brackets, rows represent targets, columns predicted classifications).

Encoding 1	FIN	PEN	ANT
FIN	1354 (70.37)	486 (25.26)	84 (4.37)
PEN	105 (4.08)	2408 (93.51)	62 (2.41)
ANT	45 (12.47)	138 (38.23)	178 (49.30)
Encoding 2	**FIN**	**PEN**	**ANT**
FIN	1692 (87.94)	194 (10.08)	38 (1.98)
PEN	148 (5.75)	2369 (92.00)	58 (2.25)
ANT	45 (12.47)	95 (26.31)	221 (61.22)
Encoding 3	**FIN**	**PEN**	**ANT**
FIN	1710 (88.88)	181 (9.41)	33 (1.71)
PEN	152 (5.90)	2379 (92.39)	44 (1.71)
ANT	53 (14.68)	84 (23.27)	224 (62.05)

antepenultimate stress (ANT) in 4.37% of the words (two remaining cells in the top row). This means that 25.26% of the words that should receive final stress actually were classified as words with penultimate stress.

If IBL performed without any misclassifications, the confusion matrix would look perfectly diagonal. But that is not the case: from the confusion matrices in Table 5 it appears that for a training set of 4000 items, the rows representing stress on the penultimate syllable are almost perfect (no exorbitant migration to other cells on the same row). For the targets FIN and ANT (respectively, the first and third rows in the matrices) this does not hold to the same extent. Moreover, there is a remarkable difference between Encoding 1 on the one hand and Encodings 2 and 3 on the other hand with respect to the misclassification of words that have FIN and ANT as their target categories. Significantly, more items arrive in the PEN category when IBL uses Encoding 1 than when Encoding 2 or Encoding 3 are used.

A third analysis that shows this overgeneralization looks as follows. We calculated Tanimoto's dichotomy coefficient (Gower 1985) for each individual target category. This statistic compares the number of words in the lexicon that have a particular target category, in this case PEN, with the predictions of IBL for those words. Thus it takes into account the proportion of agreements between the targets and the predictions. The measure is standardized by all possible patterns of agreements and disagreements. For the target category PEN the dichotomy coefficient equals .720 for Encoding 1, .815 for Encoding 2, and .827 for Encoding 3. Hence, the overgeneralization of the PEN category is reflected in the lower value of the dichotomy coefficient for Encoding 1 as compared with Encoding 2 and Encoding 3.

In this section we showed that an encoding of the input material using weight strings (Encoding 1) yields inferior results as compared with an encoding that uses a phonemic representation. This finding was substantiated both at the level of the general performance of the classifier and at the level of the individual target categories. In Encoding 1, the weight string representation was seen to find the most frequent pattern, viz. stress on the penultimate syllable, and this pattern was overgeneralized (with no recovery when more training items were used). This overgeneralization was not nearly as pronounced in Encodings 2 and 3.

A comparison of Encoding 2 (rhyme projections) and Encoding 3 (full phonemic representation) shows that in general Encoding 2 yields slightly worse results than Encoding 3, but the peak performance of both encodings does not manifest a statistically significant difference.

In the following sections we will analyze the results of the classifier in view of the metrical analysis of the Dutch data presented in Section 2.

4.3.2 Analysis of the Acquisition of General Tendencies. In Section 2.1 an overview of the stress patterns in our lexicon was provided. Three near exceptionless generalizations were pointed out. The first generalization, viz. stress is restricted to a three-syllable window from the right-hand word edge, could of course not be tested because of the format of the training material. The other two generalizations constitute a good test of how well IBL traced the main regularities governing the domain. A first test concerns the exceptionless generalization that a syllable containing a schwa is never stressed, and that words with a final schwa syllable get penultimate stress almost without exception. IBL's predictions for words with a final schwa syllable are presented in Table 6.

IBL clearly caught the generalization that if the final syllable contains a schwa, stress lands on the penultimate syllable. More than 99% of these words are classified correctly. It (over)generalized this rule also to cases in which the schwa syllable has an empty onset, in which case stress is on the antepenultimate syllable instead of the penultimate. As can be appreciated from the results, the three encodings do not differ significantly.

Table 6
Success scores for words with a final schwa syllable.

Encoding Number	Error	Correct	% Correct
1	9	1316	99.32
2	8	1317	99.40
3	8	1317	99.40

Table 7
Prediction of antepenultimate stress relative to the content of the penultimate syllable.

Structure of Penult	Lexicon	Encoding 1	Encoding 2	Encoding 3
-VV-	313	255	264	297
-VC-	7	4	6	0
-ə-	41	25	24	25
-VXC-	0	0	0	2

A second general tendency relates to the content of the penultimate syllable of words that receive antepenultimate stress. Antepenultimate stress may occur in a VV-penult but not in a VC-penultimate syllable. In Table 7 we show the number of words adhering to each pattern in the lexicon and the number of words that receive antepenultimate stress in the three encoding conditions.

IBL definitely captured this regularity in the data: in the three experimental conditions the number of ANT responses for a VC penult is extremely limited, while the number of ANT responses for a VV penult is the common case.

These findings suggest that IBL detected the strong generalizations governing the domain. In the following section we will investigate whether this also holds for the cases distinguished in the metrical analysis.

4.3.3 Analysis of Learning and Markedness. In the theoretical analysis we pointed out that approximately 80% of the data were regular according to a metrical analysis. Deviations from the regular pattern were handled by marking the deviant words in the lexicon. It was argued that only two exception features were required, viz. [−ex], for exceptions on the extrametricality condition, and LF for prespecified lexical feet. It was also pointed out that degrees of markedness followed from cumulation of these two features, yielding the following markedness scale: R(egular) < [−ex] or LF < [−ex] and LF < I(rregular). When we classify the words in our lexicon relative to the lexical marking they need and plot the results for each class, a highly illuminating picture (Figure 3) appears.

A first observation that can be made from Figure 3 is that the regular cases (R) are learned almost perfectly using the three encodings. However, there is still an important difference among the success scores of Encoding 1 (99.24%) and Encoding 2 (92.88%) and Encoding 3 (92.90%). Encoding 1 yields a significantly higher success score than the two other encodings ($p < .01$ in the χ^2-test). In other words, an encoding in terms of syllable weight proves to be almost completely predictive for handling the regular cases, even without intermediate structures such as feet and word trees.

Encodings 2 and 3 are less successful for the regular cases. However, for the marked categories, the weight string encoding (Encoding 1) does not even attain a

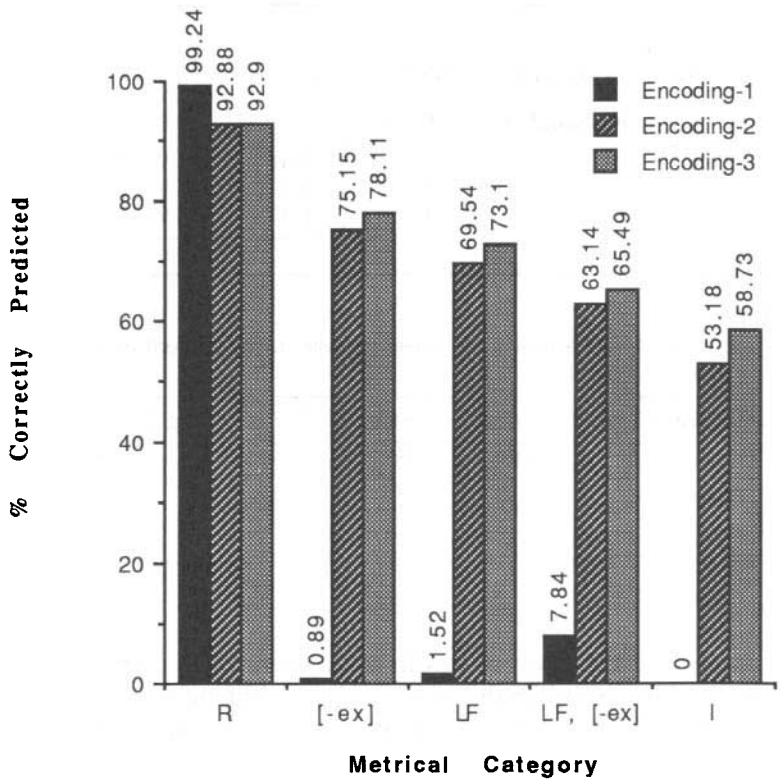


Figure 3
Success rates per 'metrical category.'

success score of 10%. Encodings 2 and 3 are far more successful in this respect. This is most impressive for words that are marked as exceptional to extrametricality ([−ex]): performance increases from 0.89% for Encoding 1 to 75.2% for Encoding 2, and even to 78.1% for Encoding 3. Taken together, the results for the marked classes of words explain the observed global performance differences between Encoding 1 and the other two encodings.

As for the comparison of Encodings 2 and 3, Figure 3 shows that Encoding 3 consistently scores higher, but none of the comparisons of the success scores yields a statistically significant difference.

An analysis of the learning results for Encodings 2 and 3 from the perspective of the metrical markedness scale reveals an interesting correspondence. The less marked a class of words is according to the metrical analysis, the better the class is learned in the experiments:

Markedness Scale:	R < [−ex] or LF < LF, [−ex] < I
Learning Performance	R < [−ex] < LF < LF, [−ex] < I[12]

12 All comparisons reveal a statistically highly significant difference, $p < .01$, as measured by the χ^2-test.

Table 8
Results for VV-final words relative to their metrical analysis.

Syllable Pattern	Stress	Marking	# Words	Encoding 2	Encoding 3
VV-VV	PEN	R	201	96.52	89.55
	FIN	[−ex], LF	64	56.25	54.69
VC-VV	PEN	R	188	98.94	97.34
	FIN	[−ex], LF	40	75.00	80.00
VXC-VV	PEN	R	17	100	100
	FIN	[−ex], LF	3	66.67	66.67
-VV-VV	PEN	R	212	75.47	81.60
	ANT	LF	124	66.94	73.39
	FIN	[−ex], LF	91	52.75	59.34
-VC-VV	PEN	R	126	92.06	96.83
	FIN	[−ex], LF	34	67.65	64.71
-VXC-VV	PEN	R	1	100	100
	FIN	[−ex], LF	2	100	100

Hence, the regular cases are learned best, while the success rate for the irregular words is lowest. In between, the words that require one single marking are learned better than those that require two markings. Thus, the markedness relations between those classes of words are reflected in the success scores. In the metrical analysis, no predictions are made about the relative degree of markedness of [−ex] versus LF. Yet in our experiments Encoding 2 and Encoding 3 agree that [−ex] words are easier to learn than LF words.

These results lead us to conclude that there is a close overall correspondence between markedness as a function of the number of lexical markings needed for particular classes of words and the learnability of those words: for unmarked classes of words, the learning algorithm reaches a superior success score than for the marked classes, and performance decreases as the number of markings increases.

Does this close correspondence between markedness in the metrical framework and learnability in the computational context also hold when we scrutinize the results for specific types of words? To examine this in detail, we will look at different types of words as was done in Tables 1 and 2.

For the ə-final words success rates for the regular pattern are 99.39% for Encoding 2 and 99.47% for Encoding 3. The handful of exceptions that require full lexical marking (I) were all wrongly classified as regular (R), so that for superlight syllables a drastic difference in performance exists between words on different ends of the markedness scale.

For words ending in superheavy syllables too, alternation exists only between R(egular) and I(rregular) patterns. The regular ones (final stress) are predicted with 96.80% accuracy for Encoding 2 and 96.35% for Encoding 3. The irregular patterns reach success scores of 55.83% for Encoding 2 and 61.67% for Encoding 3, again yielding a highly significant difference.

For words ending in light or heavy syllables, the situation is slightly more involved; here the R(egular) pattern alternates with different kinds of marked patterns, depending on the form of the prefinal syllable. In the following tables we select those types from Table 1 in which the regular pattern alternates with cases that need [−ex], LF, or a combination of both. Thus, in Table 8 and Table 9 alternations between R(egular) and I(rregular) are left out.

Table 9
Results for VC-final words relative to their metrical analysis.

Syllable Pattern	Stress	Marking	# Words	Encoding 2	Encoding 3
VV-VC	PEN	R	177	75.71	75.71
	FIN	[−ex]	134	78.36	79.10
VC-VC	PEN	R	135	77.78	79.26
	FIN	[−ex]	101	80.20	81.19
VXC-VC	PEN	R	8	50.00	62.50
	FIN	[−ex]	5	60.00	80.00
-VV-VC	ANT	R	170	67.06	63.53
	PEN	LF	64	61.19	69.84
	FIN	[−ex]	67	73.02	68.66
-VC-VC	PEN	R	36	88.24	88.23
	FIN	[−ex]	12	66.67	91.67

When we look at the data for VV-final words, the correspondence between relative markedness and learnability holds across the board. For bisyllabic words, performance for the regular case varies between 90 and 100%, whereas final stress, which needs two exception features, is predicted with success rates ranging from 55 to 80%. The differences between both categories are statistically significant in each case ($p < .01$ in the χ^2 test). For trisyllabic and longer words, the -VV-VV-type is the most interesting one, because the regular pattern (PEN) alternates with two different marked patterns, i.e. ANT, which results from a lexical foot and FIN, which needs both LF and [−ex]. For both Encoding 2 and Encoding 3 the regular pattern is learned best, followed by the one that needs a single lexical marking. The most marked pattern is the hardest to predict. For -VC-VV words, the regular case is once again learned better than the marked one. Thus, the results for individual VV-final word types corroborate the correspondence between markedness from a theoretical perspective and ease of learning for the algorithm.

For the VC-final words this correspondence does not seem to hold. The marked pattern [−ex] is predicted better than the regular pattern in most cases. This might be related to the fact that unlike with the VV-final words, the contrast between regular and marked involves only a single lexical marking. Yet, the high success rate for final stress ([−ex] words), which was already pointed out in Section 4.3.1, merits further discussion, because it seems to imply that the phonemic encodings permit the algorithm to capture relevant generalizations governing the presence of [−ex].

Closer scrutiny of the results reveals that the high performance for [−ex] words can be attributed to the fact that the algorithm has discovered subregularities in the data that are tied to segmental information and hence cannot be captured using syllable weight alone. For instance, the high success scores for final stress in VC-final words is due to a considerable extent to the fact that almost half of these words (48%) have /ɛ/ in their final syllable. Success rates for predicting final stress for these words are 92.90% for Encoding 2 and 95.74% for Encoding 3, while the success rates for this group as a whole (i.e., including those with PEN and ANT stress) are 82.82% for Encoding 2 and 84.66% for Encoding 3. The IBL algorithm also seems to have discovered the more general subregularity in the lexicon with respect to words ending in /ɛ/, viz. they almost unanimously prefer final stress (88% vs. 9.53% PEN and 2.46% ANT on a total of 325 words). This strikingly homogeneous behavior of words with /ɛ/ in their final syllable is reflected in a success rate of 88% for Encoding 2 and 89.53% for Encoding 3.

For the regular (superheavy) words in this class, the success rate was as high as 93.21% for Encoding 2 and 94.44% for Encoding 3.

While this illustrates how the segmental information in Encodings 2 and 3 enables the algorithm to learn marked patterns, these subregularities sometimes cut across the metrical classification based on syllable weight. This phenomenon can also be illustrated for other types of words: words with /e:/ in their final syllable have a strong preference for final stress (96.67%), irrespective of whether the final syllable is closed (i.e. superheavy, and hence R) or open (and hence requiring both LF and [−ex] in the theoretical analysis). The regular words are stressed with 97.62% accuracy for both encodings and the marked ones with approximately 90%, yielding a total success rate of 94.29% for Encoding 2 and 94.76% for Encoding 3.

The breadth of the ability to trace subregularities in the data based on segmental information is further illustrated by the following example: 25% of VC-final words have /u:/ in their final syllable. Of these words 48.08% have stress on the penultimate syllable and 44.23% have stress on the antepenultimate syllable. Yet the success rate for this class of words is 81.54% for Encoding 2 and 83.46% for Encoding 3, which is more than expected given the distribution of target categories. It appears (again) that the algorithm discovered finer distinctions within this set of words. A particularly striking one concerns Latinate words in /i:u:m/ (i.e., /i:/ in the penultimate syllable and /u:m/ in the final syllable). These words have antepenultimate stress in 95.24% of the cases, and a success rate of 95.24% (both encodings) for this type of words indicates that the algorithm has successfully captured this minor generalization.

In summary, we found a correspondence between markedness in the metrical framework and ease of learning by the algorithm. This correspondence was first observed on a global level, where the regular cases in the metrical analysis were learned more accurately than the marked cases for which the metrical analysis proposes lexical markings. The correspondence was also found to a large degree at the level of individual word types.

4.3.4 Summary of Results. The experiment set out to investigate the ability of IBL to acquire main stress assignment in Dutch monomorphemic words. The system was largely successful in this enterprise: its general performance attained a success score of almost 90%. The experimental findings clearly indicate that the major generalizations in the domain were captured (i) although the learning material was noisy to a considerable extent, and (ii) without using the tree-building operations deemed necessary in learning theories in the framework of metrical phonology.

In order to investigate the effects of the knowledge provided to IBL in the training examples, three encodings were used in the experiment, varying in the degree of 'theoretical bias.' The encoding incorporating the metrical notion 'weight,' as represented in the weight string encoding, was less successful overall than the encodings in terms of the actual phonemic content of the words. This finding shows that important information was lost in the abstraction of syllable weights from the phonemic content. The relative poverty of the weight string representation, which is interpreted as necessary and sufficient for stress assignment in metrical phonology, resulted in an overprediction of the most frequent pattern. The phonemic encodings, which make less or no abstraction of the segmental details, were less prone to overgeneralize the most frequently observed pattern in the input.

The performance of the algorithm shows some interesting relationships with a metrical analysis. On a general level, the success rates of IBL correlated with a markedness scale that was defined in terms of the idiosyncratic marking of words. Regular words from a metrical perspective do not require specific marking, and they were learned

very successfully. Irregular words, the other extreme on the markedness scale, showed the lowest success rates. In between these two extremes, words that require one feature are more easily learned than words that require two features. This correspondence between relative markedness and relative ease of acquisition is consistent with Gupta and Touretzky's (1994) results.

When tracing the correspondence between relative markedness and ease of learning down to the level of individual types of words, the analysis was quite successful again: for ə-final, VXC-final and VV-final words, the stress pattern of regular words is more accurately predicted than the stress pattern of marked words. Moreover, the more marked a type of word is in metrical terms, the lower the success rate for that type turns out to be.

The sole exception to this correspondence was the class of VC-final words. For these words, a marked pattern, viz. the [−ex] type, was found to be as easily (or even more easily) learned than the regular type. To explain this deviant finding, the processing of the system was traced. It turned out that the algorithm detected subregularities in the data. These subregularities could be defined in terms of characteristic segments in particular positions in the word, or clusters of segments. This is why VC-words with final stress that are marked in the metrical analysis actually turn out to be fairly regular, as judged from the learning results. 'Markedness' in metrical phonology is defined relative to an analysis in which segmental information is abstracted away in the derivation of syllable weight. But the subregularities detected by the algorithm were shown to be defined in terms of segmental information, especially vowel quality. Moreover, they cut across the metrical classification in terms of syllable weight, and, hence, markedness as conceptualized here.

The finding that, on the one hand, a metrical analysis reveals the dominant patterns in the data, but, on the other hand, does not capture important subgeneralizations in the domain, may be considered as an indication that metrical analyses should pay more attention to segmental information than is the case at present. Our research shows some directions in which such a quest can proceed.

5. Conclusion

Dresher and Kaye (1990:146) argue that "A rich and highly structured theory of UG [Universal Grammar] is otiose if the same results can be achieved by simpler means." What might these alternatives be? A possible alternative is a data-oriented one, which can be described as follows: it appears that stress patterns are sensitive to sequences of syllables and syllable weights. We could simply map strings of weighted syllables (weight strings) into sequences of stresses (stress strings). In this way a record would be kept of the stress strings associated with each weight string. This alternative was suggested by Church (1992)[13] and by Dresher and Kaye (Dresher and Kaye 1990; Dresher 1992). Dresher (1992) concludes the discussion of this alternative by stating that it is *empirically inadequate*: "It would be unable to project its grammar to assign stress to weight strings not yet encountered" (Dresher 1992: 301).

In this paper we investigated a learning device that incorporates the very simple data-driven alternative described above. The memory of an Instance-Based Learning (IBL) system is a kind of table in which representations of words are associated with

13 Church (1992), in a reaction to Dresher (1992) proposes lookup in a table of syllable weight strings (associated with their stress string) as an alternative approach. However, he glosses over the problem of ambiguity and noise, and of how to arrive at a syllable weight representation on the basis of the spelling of words.

stress assignments. However, by using simple similarity-based reasoning, the algorithm can generalize beyond the data on which it was trained.

We showed that IBL was able to acquire a considerable portion of the regularities governing the stress system of Dutch. This finding is in agreement with a similar enterprise undertaken by Gupta and Touretzky (1994), who used a simple perceptron as their data-driven approach, and shows that the tree-building operations proposed in learning theories for metrical phonology are not necessary for learning stress assignment.

We also investigated two other aspects of the learning process. First of all, in the learning experiments described in the literature thus far, only stereotyped representations of the stress patterns of languages have been used as learning material. In this study we used a lexicon of 4868 attested monomorphemes. This lexicon showed all the general characteristics of the intricate Dutch stress system, but it also contained a fair amount of noise: exceptional words from a metrical point of view as well as plain irregular cases. It was shown that IBL discovered the regularities despite the noise.

Secondly, we investigated the effect of using different representations for the training material of the learner on the learning results. The input encodings reflected the amount of 'theoretical bias' or a priori knowledge that a learner could be provided with. More specifically, a weight-string encoding is considered to be necessary and sufficient in the literature for learning a quantity-sensitive language such as Dutch. We contrasted such an encoding of our learning material with an encoding that consisted of rhyme projections, and with a plain phonemic representation (that included syllable boundaries). It turned out that the phonemic representations yield significantly better results than the encoding in terms of syllable weights. This implies that a data-driven approach to the task of acquiring main stress assignment is feasible even without the a priori knowledge incorporated in weight-strings.

Taken together, our results suggest that the representations and operations specified by metrical theory may be neither necessary nor sufficient for learning stress assignment. More specifically, information about segmental content may warrant more attention in metrical phonology. More generally, the results weaken Dresher and Kaye's (1990) argument for the necessity of a principles and parameters approach to the acquisition of stress.

Acknowledgments

We thank G. De Schutter, A. Dirksen, P. Gupta, B. MacWhinney, M. van Oostendorp, and anonymous *Computational Linguistics* reviewers for many helpful comments that led to substantial improvements of both the content and the presentation of this paper. Thanks also to Antal van den Bosch, who cooperated with us in earlier stages of this research. The research of S. Gillis and G. Durieux was supported by a research grant 'Fundamentele Menswetenschappen' (8.0034.90). This research was completed while S. Gillis was a visiting researcher at the department of Psychology of Carnegie Mellon University (Pittsburgh, PA) on a travel grant from the National Fund for Scientific Research (Belgium) and with support from CMU, which is gratefully acknowledged.

References

Aha, David W.; Kibler, Dennis; and Albert, Marc K. (1991). "Instance-based learning algorithms." *Machine Learning* 6, 37–66.

Burnage, Gavin (1990). "CELEX. A Guide for Users." Centre for Lexical Information, University of Nijmegen.

Chomsky, Noam (1981). "Principles and parameters in syntactic theory." In: *Explanation in Linguistics: The Logical Problem of Language Acquisition*, edited by Norbert Hornstein and David Lightfoot, 32–75. Longman.

Church, Kenneth W. (1992). "Comment on computational learning models for metrical phonology." In *Formal Grammar: Theory and Implementation*, edited by Robert Levine, 318–326. Oxford University Press.

Daelemans, Walter, and van den Bosch, Antal (1992). "Generalization performance

of backpropagation learning on a syllabification task." In *Connectionism And Natural Language Processing*, edited by Marc Drossaers and Anton Nijholt, 27–38. *Proceedings, Third Twente Workshop On Language Technology*, Twente, The Netherlands.

Daelemans, Walter; Gillis, Steven; Durieux, Gert; and van den Bosch, Antal (1993). "Learnability and markedness in data-driven acquisition of stress." In *Computational Phonology*, edited by T. Mark Ellison and James M. Scobbie. Edinburgh Working Papers in Cognitive Science 8, 157–178.

Derwing, Bruce L., and Skousen, Royal (1989). "Real time morphology: Symbolic rules or analogical networks." *Berkeley Linguistic Society* 15, 48–62.

Devijver, Pierre A., and Kittler, Josef (1982). *Pattern Recognition. A Statistical Approach.* Prentice-Hall.

Dresher, Elan (1992). "A learning model for a parametric theory in phonology." In *Formal Grammar: Theory and Implementation*, edited by Robert Levine, 290–317. Oxford University Press.

Dresher, Elan, and Kaye, Jonathan (1990). "A computational learning model for metrical phonology." *Cognition* 34, 137–195.

Gillis, Steven; Durieux, Gert; Daelemans, Walter; and van den Bosch, Antal (1992). "Exploring artificial learning algorithms: Learning to stress Dutch simplex words." Antwerp Papers in Linguistics 71, U.I.A., Wilrijk, Belgium.

Goldsmith, John A. (1990). *Autosegmental and Metrical Phonology.* Basil Blackwell.

Gower, J. C. (1985). "Measures of similarity, dissimilarity, and distance." In *Encyclopedia of Statistical Sciences, Vol 5*, edited by Samuel Katz and Norman L. Johnson. Wiley.

Gupta, Prahlad, and Touretzky, David S. (1994). "Connectionist models and linguistic theory: Investigations of stress systems in language." *Cognitive Science*, 18(1), 1–50.

Halle, Morris, and Vergnaud, Jean-Roger (1987). An Essay on Stress. MIT Press.

Hayes, Bruce P. (1981). *A metrical theory of stress rules.* Doctoral dissertation, Massachusetts Institute of Technology (distributed by the Indiana University Linguistics Club, Bloomington, Indiana).

Hulst, Harry and Smith, Norval (1982). *The structure of phonological representations.* Foris.

Kager, René (1985). "Cycliciteit, Klemtoon en HGI." *Spektator*, 14, 326–331.

Kager, René (1989). *A Metrical Theory of Stress and Destressing in English and Dutch.* Foris.

Kager, René, and Zonneveld, Wim (1986). "Schwa, syllables, and extrametricality in Dutch." *The Linguistic Review*, 5, 197–221.

Nosofsky, R.; Clark, S.; and Shin, H. (1989). "Rules and exemplars in categorization, identification and recognition." *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 282–304.

Nyberg, E. (1991). A non-deterministic, success-driven model of parameter setting in language acquisition. Doctoral dissertation, Carnegie Mellon University.

Quinlan, John Ross (1986). "Induction of decision trees." *Machine Learning* 1, 81–106.

Riesbeck, Christopher K., and Schank, Roger S. (1987). *Inside Case-Based Reasoning.* Erlbaum.

Rumelhart, David E.; Hinton, Geoffrey E.; and Williams, R. J. (1986). "Learning internal representations by error propagation." In *Parallel Distributed Processing: Explorations into the Microstructure of Cognition*, Vol. 2, edited by David E. Rumelhart and Jay L. McClelland, 216–271. MIT Press.

Skousen, Royal (1989). *Analogical Modeling of Language.* Kluwer Academic Publishers.

Smith, Edward E., and Medin, Douglas L. (1981). *Categories and Concepts.* Harvard University Press.

Stanfill, Craig, and Waltz, David L. (1986). "Toward memory-based reasoning." *Communications of the ACM*, 29, 1213–1228.

Trommelen, Mieke, and Zonneveld, Wim (1989). *Klemtoon en Metrische Fonologie.* Coutinho.

Trommelen, Mieke, and Zonneveld, Wim (1990). "Stress in English and Dutch: A comparison." Dutch Working Papers in English Language and Linguistics 17, R.U.L., Leiden, The Netherlands.

Trommelen, Mieke (1983). *The Syllable in Dutch.* Foris.

Trommelen, Mieke (1991). "Dutch word stress assignment: Extrametricality and feet." In *The Berkeley Conference on Dutch Linguistics 1989. Issues and Controversies, Old and New*, edited by Thomas F. Shannon and Johan P. Snapper, 157–172. U.P.A., Lanham.

van der Hulst, Harry, and Smith, Norval (1982). *The Structure of Phonological Representations.* Dordrecht: Foris. Two volumes.

van der Hulst, Harry (1984). *Syllable Structure and Stress in Dutch.* Dordrecht: Foris.

Van Marle, Jaap (1978). "The stress pattern

of Dutch simplex words: A first approximation." In *Studies in Dutch Phonology*, edited by Wim Zonneveld, F. Van Coetsem, and O. W. Robinson, 79–121. Martinus Nijhoff.

Visch, Ellis, and Kager, René (1984). "Syllable weight and Dutch word stress." In *Linguistics in the Netherlands 1984*, edited by Hans Bennis, and W. U. S. van Lessen Kloeke, 197–205. Foris.

Weiss, Sholom M., and Kulikowski, Casimir A. (1991). *Computer Systems That Learn*. Morgan Kaufmann.

Zonneveld, Wim (1993). "Schwa, superheavies, stress and syllables in Dutch." *The Linguistic Review* 10, 61–110.