Proceedings 15th conference of the cognitive science society, 452-457, 1993

Learnability and Markedness: Dutch Stress Assignment*

Steven Gillis†, Walter Daelemans‡, Gert Durieux†, Antal van den Bosch‡
†National Fund for Scientific Research (Belgium)
Department of Linguistics, University of Antwerp
Universiteitsplein 1
B-2610 Wilrijk
gillis@reks.uia.ac.be, durieux@reks.uia.ac.be
‡Institute for Language Technology and AI (ITK)
Tilburg University
P.O. Box 90153
NL-5000 LE Tilburg
walter@kub.nl, antalb@kub.nl

Abstract

This paper investigates the computational grounding of learning theories developed within a metrical phonology approach to stress assignment. In current research the Principles and Parameters approach to learning stress is pervasive. We point out some inherent problems associated with this approach in learning the stress system of Dutch. The paper focuses on two specific aspects of the learning task: we empirically investigate the effect of input encodings on learnability, and we examine the possibility of a data-oriented approach as an alternative to the Principles and Parameters approach. We show that a data-oriented similarity-based machine learning technique (Instance-Based Learning), working on phonemic input encodings is able to learn metrical phonology abstractions based on concepts like syllable weight, and that in addition, it is able to extract generalizations which cannot be expressed within a metrical framework.

Introduction

Machine learning of metrical phenomena is an interesting domain for exploring the potential of particular machine learning techniques, and more generally, to study the role Machine Learning can play in theory formation (the *computational grounding* of a learning theory). Not only are available a solid, relatively independent, theoretical framework and elaborate descriptions of the linguistic data, it is also the case that metrical phenomena exhibit the combination of generalization, subregularity and exceptions which is typical of linguistic phenomena in general.

Machine learning of metrical phenomena

Recently, several computational learning models that specifically address the problem of how to learn the regularities of stress assignment have been proposed: Gupta & Touretzky (1991), Dresher & Kaye (1990), Nyberg (1992). They all approach the learning problem from the angle of the 'principles and parameters' theory (Chomsky 1981). In this approach the learner comes to the task of language learning equipped with a priori knowledge incorporated in a universal grammar that constrains him to entertain only useful generalizations. It is assumed that universal grammar specifies a number of parameters relevant to the metrical domain (see Dresher & Kaye, 1990). The computational models add a *learning theory* to the linguistic notion of universal grammar. This theory specifies what aspects of the data are relevant to each parameter, and it also determines how the data processed by the learner are to be used to set the values of the parameters. Common to the systems referred to is that they try to fix the values of parameters relevant to the metrical domain.

Stress assignment in Dutch

In order to introduce the parameter setting problems in the P&P approach for the case of Dutch, a short presentation of some basic facts about the stress system of Dutch appears to be in order. The most straightforward way to present stress assignment in Dutch is by reviewing the settings of the relevant metrical parameters (see Dresher & Kaye, 1990, Trommelen & Zonneveld, 1990):

*The research of Steven Gillis and Gert Durieux was supported by a Research Grant S 2/5 - CL.D98 from the National Fund for Scientific Research (Belgium), and a research grant "Fundamentele Menswetenschappen" (8.0034.90). We are grateful to Georges De Schutter and Arthur Dirksen for useful comments and discussion.

P1	The word-tree is strong on the [Left/ Right]	Right
P2	Feet are [Binary/Unbounded]	Binary
P3	Feet are built from the [Left/Right]	Right
P4	Feet are strong on the [Left/Right]	Left
P5	Feet are quantity sensitive [Yes/No]	Yes
P6	Feet are quantity sensitive to the [Rhyme/Nucleus]	Rhyme
P8A	There is an extrametrical syllable [No/Yes]	Yes
P8	It is extrametrical to the [Left/Right]	Right

From these settings, it can be inferred that the unmarked case in Dutch is stress on the penultimate syllable. Taking the extrametricality parameters into account (parameters 8A and 8) the possibility is created for the antepenultimate syllable to be stressed[1]. Stress on the last syllable is, according to this theory, fairly marked, except for super heavy syllables which are stressed almost without exception. Super heavy syllables are not subjected to the extrametricality condition.

Deviations from the unmarked case are handled as follows:

- **Lexical Feet [F]**. The mechanism of idiosyncratically assigning a lexical foot stipulates that the syllable marked as constituting a lexically prespecified (monosyllabic) foot behaves as an exception to regular foot construction.
- **Exceptions to the extrametricality rule [-ex]**. This mechanism indicates that words marked as [-ex] are to be withdrawn from the regular application of the extrametricality rule. The aim of this marking is to attract stress to a final syllable that would be extrametrical in the regular case.
- **Lexical feet in conjunction with exceptions to extrametricality [F], [-ex]**. The third mechanism combines the two preceding ones: it marks a final syllable so that it is assigned a monosyllabic foot, and subsequently this syllable is withdrawn from the regular application of the extrametricality rule by a [-ex] marking. The three exception mechanisms have in common that the relevant words have to receive a marking in the lexicon. This also holds for those words that can still not be satisfactorily treated by the mechanisms discussed. Their stress pattern is indicated in the lexicon as well.

[1]In Dutch, every VX-rhyme is considered extrametrical, where X stands for V or C. This condition results in the extrametricality of VV- and VC-rhymes, but is not stretched further to include super heavy syllables (which can only occur in word final position). Furthermore, Dutch is fairly idiosyncratic in the sense that extrametricality applies after foot formation has taken place, a phenomenon called 'late'-extrametricality.

Table 1: Stress patterns in Dutch words with light and heavy syllables

Type	Stress Pattern		
	Final Stress	Penult. Stress	Antepen. Stress
VV-VV-VV	[-ex][F]	R	[F]
VX-VC-VV	[-ex][F]	R	I
VX-VC-VC	[-ex]	R	I
VX-VV-VC	[-ex]	[F]	R

In sum, we have five cases according to the metrical analysis (between brackets their frequency in a lexicon that will be described in more detail in the next section): (i) the regular (R), unmarked case (80.44%); (ii) a mechanism that intrudes into foot formation: [F] (3.86%); (iii) a mechanism that affects word-tree formation: [-ex] (7.15%); (iv) a combination of (ii) and (iii) (5.38%); (v) the irregular cases (I) (3.16%). These five possibilities can be scaled according to their *markedness*: the regular case (i) is of course the least, the irregular case (v) the most marked. In-between these extremes, possibilities (ii) and (iii) are less marked than (iv). This scaling can also be performed at a more fine-grained level. In Dutch, words with three or more syllables can receive stress on any of the last three syllables. This means that even words that have equal weights in their last three syllables, can nevertheless exhibit the three possible stress patterns attested in the language. The most interesting types are displayed in Table 1. It is indicated how they are treated in the metrical framework.

The parameter setting problem

A parametric approach that aims at universal validity will eventually have to deal with the irregular, exceptional, and language specific details of the linguistic system. At present this appears to be a problem. For instance, Dresher & Kaye (1990) explicitly require that the input be completely transparant. They dedicate a specialized module to determining if there exist obvious conflicts (such as the ones illustrated above for Dutch). Eventually, a brute force learner is invoked to deal with similar input.

They also indicate that the set of parameters will undoubtedly have to be extended (see also Gupta & Touretzky, 1991). But keeping the present set of parameters as sufficient, for the sake of the argument, a number of serious problems turn up when we try to analyze how a learner of Dutch might fix the values of the parameters. Two examples may suffice to illustrate the point. Parameter 6 relates to quantity sensitivity, and more specifically determines if a language is quantity sensitive to the rhyme or to the nucleus. If the former is the case, closed syllables and long nuclei behave similarly with respect to stress, while in the latter case only branching nuclei are heavy. It is not clear how these cues for Parameter P6 can be used in Dutch where closed syllables do indeed behave as open syllables with long vowels but this is only so for heavy closed syllables and not for super heavy ones.

Another problem arises with respect to the extrametricality parameters 8A and 8. Dresher & Kaye (1990: 189) point out that extrametricality is a difficult problem since the cue "(...) presence of stress at the left or right edge of a word is enough, in this system of parameters, to rule out extrametricality at the edge." But lack of stressed peripheral syllables is not a sufficient condition for extrametricality. In the case of Dutch there is a firm number of words exhibiting final stress (in our lexicon of 4868 polysyllabic monomorphemes: 39.59%). Thus how can the learner determine that for Dutch a parameter setting amounting to right extrametricality is appropriate given a huge number of words with final stress? Moreover, the theory should provide a way to disentangle the cues for setting parameter 8A (extrametricality) and parameter 6 (quantity sensitivity) since a branching rhyme is subject to extrametricality except for super heavy syllables (with either a branching nucleus or a branching coda). If such a construction could be found it would account for 68.35% of the cases with final stress. For the remaining words with final stress the theory should find a way to discover the application of exception mechanisms, viz. [-ex] (18.06% of words with final stress), and [F][-ex] (13.60% of words with final stress).

Experiment

In the light of the above problems, we investigated the learnability of Dutch stress assignment in a machine learning experiment. Our data consisted of 4868 polysyllabic monomorphematic Dutch words. The lexicon was extracted from the CELEX lexical database[2]. Only unambiguous monomorphemes were selected and proper nouns were withdrawn from the dataset. As such it constitutes a representative sample of the monomorphemes of the language.

Method and data coding

The experiment was performed using the leaving-one-out method: each item in the dataset is used in turn as the test item, with the remainder of the dataset as training set. We therefore get as many simulations as there are items in the dataset. This computationally very costly method has as its major advantage that it provides the best possible estimate of the true error rate of a learning algorithm (Weiss & Kulikowski, 1991).

The data were encoded (i) as strings of syllable weights of the last three syllables of the word (encoding-1), and (ii) using the phonemic information contained in the rhyme projections of the last three syllables (encoding-2). For instance, the word *nirvana* was encoded as follows:

	Encoding-1	Encoding-2	
Syllable	Weight	Nucleus	Coda
Antepenultimate	3	I	r
Penultimate	2	a	-
Final	2	a	-

The learning algorithm: Instance-Based Learning

Instance-based learning (IBL, Aha et al., 1991) is a framework and methodology for incremental supervised machine learning. 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 (MBR, Stanfill and Waltz, 1989) and Case-Based Reasoning (CBR, Riesbeck and Schank, 1989) the hypothesis that much of intelligent behaviour 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). As such IBL shares an emphasis on 'analogy' in language use with Skousen's theory of Analogical Modeling (Skousen, 1989)[3].

The operation of the basic algorithm is quite simple: for each pattern to be assigned a category (test item), it is checked whether this pattern has been encountered in the training set earlier. If this is the case, the category of the training item is assigned to the new item (or the category most often associated with the training item in case of ambiguous patterns). If the test item has not yet been encountered, its similarity to all items kept in memory is computed, and a category is assigned based on the category of the most similar item(s). The performance of an IBL classifier crucially depends on the selection of training items to be kept in memory,

[2]CELEX contains 130,778 lemmas and 399,816 wordforms. It was compiled on the basis of the INL corpus of present-day Dutch (more than 42 million words in a variety of text types).

[3]The same experiments were also performed with the Analogical Modeling algorithm (Skousen, 1989) and with the Backpropagation of Errors algorithm (Rumelhart et al., 1986). A comparative analysis of the results of the three algorithms is reported in Gillis et al. (1992).

and the similarity metric used. In the experiment reported in this paper, all training items were "remembered". We only experimented with the similarity metric. We extended the metric proposed by Aha et al. (1991) with a technique for 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).

Results

The algorithm attains an overall success rate of 81.2% for encoding-1 and 87.6% for encoding-2. Specified to the level of individual target categories, it appears that stress on the penultimate syllable can be more efficiently predicted than stress on the final syllable. Stress on the antepenultimate syllable is fairly difficult to predict.

	Encoding-1	Encoding-2
Final	70.5	86.0
Penultimate	93.7	91.9
Antepenultimate	49.6	64.5
Total	81.2	87.6

A comparison of the results indicates that IBL takes advantage of the details provided in encoding-2: in general, an encoding in terms of weight strings does not lead to better results than an encoding in which the nucleus and the coda are fully specified. The difference between the results for the two encodings is statistically significant ($p < .0001$).

Specified at the level of the individual target categories, encoding-1 yields better results for the unmarked case (stress on the penultimate syllable), while for the other two target categories, the second encoding scheme yields better results. It appears that the regularities governing stress in the latter cases require information present in encoding-2, and absent in encoding-1. Thus the question arises which generalizations within the domain are captured by training the systems with the two encodings?

Weight strings versus rhyme projections. The fact that a number of general characteristics of stress assignment can be captured given the weight string encoding can be shown by scrutinizing strong generalizations within the domain that can be formulated in terms of syllable weight: (i) super heavy final syllables receive stress almost without exception (in our lexicon: 1015 words with final super heavy syllable receive final stress, while in only 56 words stress is not final); (ii) super light syllables can never be stressed and are almost without exception preceded by a stressed syllable (1316 words in our lexicon have prefinal stress when the final syllable is super light and only 11 have stress on the antepenultimate).

	Encoding-1	Encoding-2
Super Heavy		
-VVC	1015 (94.77)	1008 (94.12)
-VCC	302 (82.51)	301 (82.24)
total	1317 (91.65)	1309 (91.09)
Super Light	1316 (99.17)	1317 (99.25)

It can readily be inferred from these results that the generalizations that can be formulated in terms of syllable weight are well captured. The phonemic encoding (encoding-2) does not yield highly superior results, on the contrary, with respect to super heavy syllables in word-final position, slightly worse results are found. (None of the differences are significant at the 5% level or below.)

For light and heavy syllables less stringent generalizations are discovered by the algorithm. Light and heavy final syllables are considered to be extrametrical in the current account, thus a prefinal stress pattern is to be expected. This expectation is only realistic, however, if only extrametricality is playing a determining role. That is not the case: for VV-final words only 65.15% actually receive penultimate stress, and for VC-final words only 43.69%. These figures sharply contrast with those for the super heavy and the super light final syllables, for which the algorithm found satisfactory generalizations. The results in the following table show that similar generalizations were out of reach for the light and the heavy final syllables.

	Encoding-1	Encoding-2
light -VV	773 (67.51)	928 (81.05)
heavy -VC	548 (57.14)	708 (73.83)

In comparison with the success scores for the super heavy and super light final syllables, the success scores of the light and heavy ones are inferior. Even for the phonemic encoding, a success rate of 80% can hardly be reached. It can be noted that the success scores for the encoding in weight strings are below the phonemic encoding (differences are significant at $p < .0001$).

Exception mechanisms and markedness. Several exception mechanisms have been invoked to account for the apparent lexical diffusion that appears to govern the words with final light and heavy syllables. When we code the results of our experiment according to those categories a highly illuminating picture occurs (Figure 1).

1. The regular cases are fairly well predicted when the algorithm is trained with a weight string encoding (99.59%). This encoding is not able to deal with the lexically marked words: the success scores for the four types of exceptions does hardly reach 10%. The results for the regular category are significantly ($p < .01$) better for the weight string encoding than for the phonemic encoding of the rhyme.

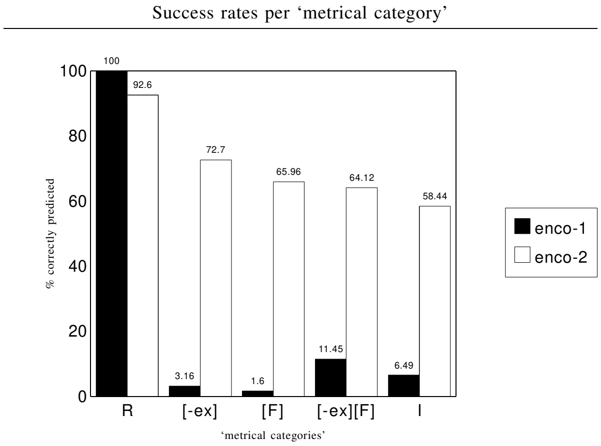

Figure 1: Performance relative to Metrical Category

2. The exceptional cases reach a fairly acceptable level of accuracy when the algorithm is trained with a phonemic encoding of the rhymes of the three last syllables. All differences between encoding-1 and encoding-2 reach significance at the 1% level or below.

3. A comparison of the learning results for the phonemic encoding with the 'markedness'-scale presented within the metrical framework, immediately reveals that there is a remarkable correspondence between the two: the more marked a category from a metrical point of view, the lower the success rate of that category in the learning experiments. Hence, the regular cases are fairly well learned, and irregular cases show a poor performance. With respect to the exception mechanisms in-between these two extremes, marking of an exception with respect to extrametricality ([-ex]) and the marking of a monosyllabic lexical foot ([F]) have better scores than the category that combines the two features. Thus, the markedness relations between these exception mechanisms are reflected in lower success scores.

These results lead us to the conclusion that there is a close correspondence between markedness in terms of exception mechanisms invoked for particular classes of words and the learnability of those words: for unmarked classes of words the learning algorithms reach a superior success score in comparison to the more marked classes.

Does this close correspondence between markedness in the metrical framework and learnability in the computational context also hold when we consider the results for specific types of words? In Table 2 the information from Table 1 is repeated and the learning results for IBL are added.

Table 2: Success scores for words with light and heavy syllables

	Stress Pattern		
Type	Final Stress	Penult. Stress	Antepen. Stress
VV-VV-VV	[-ex][F]	R	[F]
IBL	60.00	83.45	77.78
VX-VC-VV	[-ex][F]	R	I
IBL	65.63	91.06	0.00
VX-VC-VC	[-ex]	R	I
IBL	83.33	80.65	33.33
VX-VV-VC	[-ex]	[F]	R
IBL	67.16	73.33	65.29

At first sight, relative markedness from a metrical point of view appears to be a good predictor of the success scores of the learning algorithms. Take the VV-VV-VV words as an example. The regularly stressed type (stress on the penultimate syllable) has, by far, the best success score. A somewhat lower score is obtained for the words with antepenultimate stress. These words are more marked than the regulars: they need a single exception feature. Final stress is obtained for words with two features, this category, the most marked of the three, has the worst score. Thus metrical markedness is reflected in the success score of the algorithm. A similar finding holds for the VX-VC-VV words: R > [-ex][F] > I.

The relationship between markedness and success scores does not seem to be as strong when we consider the two bottom rows of Table 2. For VX-VC-VC words, the irregular antepenultimate stress, the most marked category, is very poorly predicted. But the regular penultimate stress is not predicted better than the more marked final stress. In the bottom row a similar situation is found: the regular case is less well predicted than the more marked ones. Although these results appear to contradict the relationship between markedness in metrical terms and success scores of the algorithms, this contradiction can be explained by a closer analysis of the learning patterns.

When we analyze the results for the individual patterns of words, IBL appears to have discovered subregularities in the data that are not (even, cannot) be accounted for in the metrical framework. Indeed, in the latter, syllables are used in the analysis as far as their weights are concerned. The identity of individual vowels and consonants is not taken into account in the constructions of metrical trees. And hence, for the word types considered, important subclasses of words that behave homogeneously cannot be identified. But given the phonemic encoding used in the learning experiment, the algorithm quite successfully traces these subregularities. For instance, the high success scores for final stress in words with a VX-

VV-VC pattern is partly due to the fact that almost half of these words (48%) have /E/ in their final syllable. They are successfully stressed (93.51%). IBL seems to have discovered the more general subregularity in the lexicon with respect to these words, viz. words in /E/ almost unanimously prefer final stress (94.48% final stress, 5.22% penultimate stress, and 0.3% antepenultimate stress on a total of 326 words). This outspoken homogeneous behaviour of words with /E/ in their final syllable is reflected in the success score: 88.34%. The regular words with final stress were even more accurately stressed by IBL (95.47%).

The ability to trace subregularities in the data and the breadth of that ability is further illustrated in the following example: 25% of the VC final words have /U/ in their final syllable. This category of words is an almost perfect example of lexical diffusion: 48.08% have penultimate stress and 44.23% antepenultimate stress. IBL appears to have made finer distinctions within this set of words: (Latin) words with /i/ in the prefinal syllable and /Um/ in the final syllable act as a fairly homogeneous category with respect to stress (95.24% of these words have antepenultimate stress and 4,76% penultimate stress). Hence the success score equals this proportion: 96.43% of the words are correctly stressed.

These results lead us to the conclusion that the correspondence between markedness in the metrical framework and ease of learning by the algorithm also holds at the level of individual types of words, but at this level the correspondence is not across the board. However, apparent exceptions can be accounted for by fact that the algorithm traced subregularities in the data that cannot be captured using the less fine-grained weight strings used in the metrical framework.

Conclusion

In this paper we have shown that a data-oriented similarity-based machine learning technique is able to master the task of stress assignment. Moreover it was shown that IBL working on phonemic input encodings yields superior results than when working on weight strings. This means that (i) at least for this task, the a priori knowledge required in a Principles and Parameters approach does not appear to be necessary, and (ii) the problems the P&P approach runs into, e.g., with respect to establishing extrametricality, are circumvented.

A second major finding relates to metrical phonology: our experiment shows that markedness in such a framework corresponds to ease of learning by an artificial learning algorithm. It was found that the more marked stress patterns are less accurately learned. This correspondence between metrical markedness and ease of learning was established on a global level as well as on the level of individual categories of words. Our work therefore adds computational grounding to the theory of metrical phonology as applied to Dutch. In addition, learnability of marked (exceptional) word types could be accounted for by the fact that the similarity-based algorithm extracts subregularities from the data that cannot be captured using the machinery of metrical phonology, which explains the superior performance with a phonemic encoding in comparison to a metrical, less fine-grained encoding in terms of weight strings.

References

Aha, D., Kibler, D. and Albert, M. 1991. Instance-Based Learning Algorithms. *Machine Learning* 6:37-66.

Chomsky, N. 1981. Principles and parameters in syntactic theory. In Hornstein, N. and Lighfoot, D. eds. *Explanations in linguistics: The logical problem of language acquisition.* London: Longman. 32-75.

Dresher, E. and Kaye, J. 1990. A computational learning model for metrical phonology. *Cognition* 34:137-195.

Gillis, S., W. Daelemans, G. Durieux, and A. van den Bosch. 1992. Exploring Artificial Learning Algorithms. *APIL* 71.

Gupta, P. and Touretzky, D. 1991. Connectionist models and linguistic theory: Investigations of stress systems in language. Unpublished ms.

Nyberg, E. 1992. A non-deterministic, success-driven model of parameter setting in language acquisition. Unpublished PhD, Carnegie Mellon University.

Quinlan, J. R. 1986. Induction Of Decision Trees. *Machine Learning* 1:81-106.

Riesbeck, C. K. and Schank, R.S. 1987. *Inside Case-Based Reasoning.* Hillsdale: Erlbaum.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. 1986. Learning Internal Representations by Error Propagation. In D.E. Rumelhart and J.L. McClelland (eds.), *Parallel Distributed Processing: Vol. 2.* Cambridge, MA: MIT Press.

Skousen, R. 1989. *Analogical Modeling of Language.* Kluwer: Dordrecht.

Stanfill, C. and Waltz, D.L. 1986. Toward Memory-based Reasoning. *Communications of the ACM* 29:1213-1228.

Trommelen, M. and Zonneveld, W. 1990. Stress In English And Dutch: A Comparison. DWPELL 17.

Weiss, S. and Kulikowski, C. 1991. *Computer Systems That Learn.* San Mateo: Morgan Kaufmann.