Machine Learning of Natural Language:

An Empiricist Alternative to Nativist Theory

Walter Daelemans

Steven Gillis

Gert Durieux

University of Antwerp - UIA

Address for Correspondence:

University of Antwerp -UIA
Department of Linguistics - GER
Universiteitsplein 1
B-2610 Wilrijk

Tel.: +32 3 820 27 66
Fax: +32 3 820 22 44
e-mail: gillis@reks.uia.ac.be

Machine Learning of Natural Language: An Empiricist Alternative to Nativist Theory[*]

0. INTRODUCTION

Current theorizing about language acquisition heavily relies on the idea of a priori knowledge. Linguists and psycholinguists adhering to the Chomskyan theory of Universal Grammar accept the assumption that in order to learn a natural language a substantial amount of prewired knowledge about the content and structure of language is required. This 'nativist' view has recently been implemented in computational models of language acquisition.

In this paper we first discuss the linguistic theory of Universal Grammar and a computational model of the acquisition of a subsystem, viz. metrical phonology. The structure and assumptions of the most articulated system of stress acquisition (Dresher & Kaye 1990) are critically investigated. It is shown on theory internal grounds that the nativist approach might well be supplemented by an empiricist alternative.

In two experiments using data-driven, empiricist algorithms the feasibility of a similar approach is investigated. It will be shown that (i) the algorithms reach superior results when trained with data that are not coded in a theory impregnated way (in contrast with the predictions of the 'nativist' approach) and (ii) the algorithms attain (at least) comparable results to the 'nativist' counterpart even when the exact match with memorized items is eliminated.

1. LEARNING A NATURAL LANGUAGE: THE NATIVIST VIEW

The problem of language acquisition can be quite conveniently described as follows: given a finite set of input data a human is able to extract the rules of the grammar that generate the input data, or in other words, the rules that describe the structure of those data. Given the grammatical rules, it is possible to 'understand' and/or to 'produce' any/every grammatical string of the language and to determine if a string can be accounted for by the rules of the grammar. For instance, every speaker of Dutch is able to determine that *'kliem'* is a perfectly acceptable word though it happens to be non-existent. *'Lkiemjk'*, however, exhibits a sound structure that is not acceptable: the phonotactic rules of Dutch appear to stipulate for instance that

[*] This research 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) to S. Gillis.

'kl' is possible at the beginning of a word while *'lk'* is not a possible onset of a word. Even though *'kliem'* does not (yet) exist in the language, speakers of Dutch will agree that *'kliem-pje'* is the grammatical diminutive form of *'kliem'*, and that sentence (1) is perfectly grammatical provided that *'kliem'* is taken to be a noun or they can construct a sentence like (2) if instructed that *'kliem'* is a verb.

(1) Mijn *kliem* is gisteren vertrokken. (*My kliem left yesterday*)
(2) Ik *kliem* er maar wat op los. (*I just kliem around a bit.*)

These examples indicate that a tacit result of language acquisition is a grammar of the language at various levels: the level of sounds, words, sentences, etc. Although the average language user is not able to inspect the rules of the grammar, he uses them in understanding and producing utterances, and even in making (meta-linguistic) statements about the acceptability or grammaticality of linguistic objects.

The examples also show that the input of the learning process does not consist of that grammatical knowledge: sentence (2) does not explicitly contain the information "*kliem is a verb*". We can infer that 'kliem' is a verb because according to the syntax of the language, 'kliem' occupies the position of a verb in Dutch sentences. From the point of view of the learner this turns out to be a serious problem: if the learner knows either the structural characteristics of sentence (1) he can infer the syntactic class of *'kliem'* or if he knows the syntactic class of *'kliem'* he has valuable information for analyzing the syntactic structure of the sentence. But if he has neither piece of information, it is unclear how he can eventually crack the syntactic code.

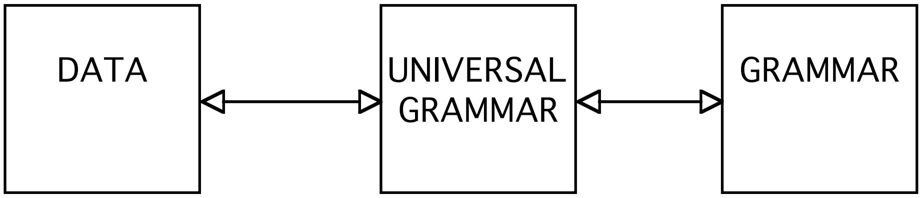
A nativist solution of this dilemma proposed in the Chomskyan tradition (Chomsky 1981, Hyams 1986) is to assume that if the necessary information is not available in the input (the 'poverty of the stimulus' argument), the learner must be endowed with knowledge that enables him to bootstrap himself into the linguistic system. Thus if the learner has innate or prewired knowledge about verbs, for instance, he can use that knowledge to learn the syntactic structures of the language in which verbs play a crucial organizing role.

What does the innate knowledge consist of? If particular knowledge is innately given (as part of a genetically determined program), it should relate to universal principles which determine the form of any human language. One of the basic insights generative grammar offers from a typological point of view is precisely that with regard to fundamental properties and organizational principles, human languages do not exhibit infinite and arbitrary variation. Instead a relatively small set of structural dimensions (or parameters) defines the possible structural variations across human languages. Each parameter represents two (or more) possible forms that a supposedly innate principle can take. In this way the set op parameters constitutes a Universal Grammar, a collection of constraints on the possible structure of a language.

The problem of language acquisition now amounts to the following: Given the data of a language and a grammar of that language, a learner learns the grammar from the data by setting the parameters specified in Universal Grammar to their language specific settings. Thus parameter setting greatly simplifies the learner's job: the learner has to attend to the right kind of evidence. On encountering that evidence (the 'cues' or 'triggers'), the relevant parameter is fixed and the learner can draw a whole series of conclusions about the nature of the target language. And by consecutively fixing parameters, the learner constructs a path through the space of possible grammars, he constructs the path that represents the grammar of the target language as an instantiation of a path in 'grammar space'.

A convenient and conventional way to represent the problem of language acquisition in this context is shown in Figure 1. Given the data of a language (D) and a grammar (G) of that language, the learner acquires G via Universal Grammar, a set of innate and universal cognitive principles.

Figure 1: Representation of Language Learning



2. An example of the nativist view

A domain in which the UG-approach is well developed concerns rhythmical phenomena, especially stress assignment. In every language of the world the syllables of a word are not equally prominent: particular syllables in a word are more prominent than others. For instance, in the Dutch word *'marmelade'* the relative prominence of the syllables can be indicated as in (3)

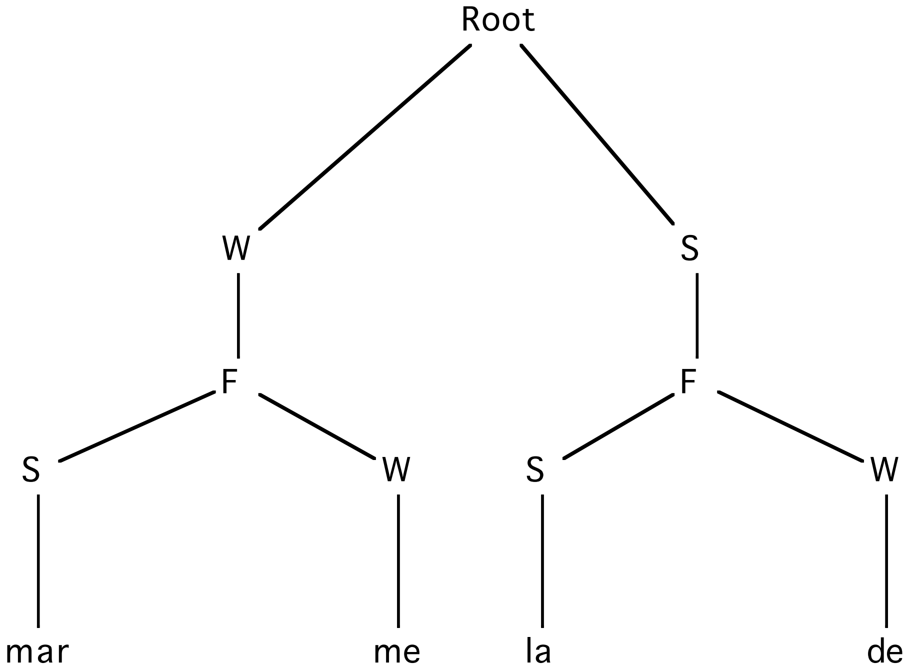
(3)

mar	me	la	de
1	0	2	0

where prominence ranges from 2, the most prominent syllable, to 0 the less prominent syllables. Intuitively it should be clear that syllables can be assigned a metrical weight: syllables containing a schwa are never stressed, i.e., relative to other syllables they are always less prominent. These syllables are said to be super light. Syllables with much more phonetic material are much stronger: in Dutch syllables with a long vowel followed by a consonant (e.g., *'materiaal''*) or a short vowel followed by two consonants (e.g., *'experiment''*) are stressed almost without exception. These syllables are said to be super heavy.

Metrical phonology is the linguistic theory that deals with phenomena such as word stress. In that framework metrical trees are drawn that capture the regularities of the rhythmical pattern of words. Take *'marmelade'* again, the stress pattern exhibits a regular alternation of more and less prominent syllables. This structure is depicted in the tree in Figure 2 by grouping syllables into rhythmical feet (F) in which the left branch in a foot is more prominent than the right one. Feet are grouped into a word structure (Root) and it appears that of the more prominent syllables, the prominent syllable of the rightmost foot is always the most prominent syllable of the word in Dutch. Thus, feet are labelled S (strong) and W (weak) and in Dutch the rightmost foot is labelled S, all other feet are W.

Figure 2: Example of a metrical tree.



In order to arrive at the syllable that carries main stress (the most prominent syllable of the word) one simply follows the S-branches in the tree: a path that consists exclusively of S's leads to the syllable with main stress.

Of course word stress in Dutch is more complicated than the structure represented in Figure 2[1], but the point is that the stress system of Dutch is fairly regular. The rules can be captured in a grammar that generates trees that look similar to the one shown above.

Metrical theory is not just a representational device tailored to the rhythmical patterns of a single language. Metrical theory is also a theory about the possible stress systems in the languages of the world. Typological studies revealed that particular stress patterns turn up time and again, while other patterns are practically non-existent. In other words natural languages appear to be very selective with respect to the stress patterns they allow: out of the theoretically possible stress systems, a well circumscribed subset actually occurs. That subset of possible and actually occurring stress systems can be defined by eleven parameters (in the right hand column the settings for Dutch):

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 [Rime/Nucleus]	Rime
P7	A strong branch of a foot must itself branch [No/Yes]	Yes
P8^{A}	There is an extrametrical syllable [No/Yes]	Yes
P8	It is extrametrical to the [Left/Right]	Right
P9	A weak foot is defooted in clash [No/Yes]	/
P10	Feet are noniterative [No/Yes]	/

The details of what these parameters stand for need not concern us here, however it should be clear that the theory makes nontrivial claims about the number and the nature of possible stress systems:

- the stress system of any single of the world's languages can be described in terms of a setting of these eleven parameters;
- the eleven binary parameters define 2^{11} (= 2048) possible stress systems, and if we take into account the interdependencies between the parameters we are left with 216 distinct stress systems. In comparison: if

[1] Take for instance words with three syllables in which each syllable ends in a long vowel. All these words do not have the same stress pattern: some have stress on the final syllable (e.g., *'paraplu''*) , some on the penultimate syllable (e.g., *'pijama''*) and still others on the antepenultimate syllable (e.g., *'Panama''*). A state-of-the-art analysis can be found in Trommelen & Zonneveld (1990).

we limit the vocabulary of stress strings to 2 and 0, the solution space consists of 4^{16} ($\cong$ 4.3 x 10^9) possibilities for strings of four syllables (Dresher 1992).

In keeping with the nativist theory sketched in the previous section, it is posited that Universal Grammar contains the parameters that control the construction of metrical structure. Consequently, in the acquisition process the learner has to figure out how to set the parameters for the language (s)he is exposed to.

3. THE LEARNING THEORY

In the previous sections we have outlined the general framework of a nativist theory of language acquisition, viz. the Universal Grammar approach. The approach was illustrated in the domain of metrical phenomena. The gist of the argument so far is that in a nativist approach Universal Grammar mediates in the acquisition process. It provides a number of parameters that relate to the main theoretical concepts and organizing principles of the linguistic theory. However neither the linguistic theory nor the conceptualization of Universal Grammar as it stands point out how the acquisition process actually proceeds. For instance, in the discussion of the *'kliem'* example in the first section we pointed out that the learner's job would be considerably simplified if he had some knowledge about *verbs* and *nouns* in the language. And indeed it is argued that syntactic categories such as Noun and Verb are part of the learner's innate luggage (Pinker 1984, 1989, Hyams 1986). But the question remains: even if the learner knows there are nouns and verbs 'out there' how can he find them? Or, even if the first metrical parameter states that a word tree is strong to the left or the right, how does the learner actually figure out what the setting of the parameter is? What evidence does he take into account? What are the cues he looks for? What are the triggers for fixing the value of the parameter? And, eventually, where do those cues or triggers come from?

To our knowledge the question about the origins of the cues remains unanswered in the literature. However it is hard to imagine how a learning theory that incorporates parameters and requires a principled relationship between its cues and the parameters of Universal Grammar (Dresher & Kaye 1990) can find another solution than incorporating the cues into Universal Grammar next to the parameters they relate to.

The use of cues or triggers is left implicit in the (psycho-)linguistic literature - the exact nature of parameter setting is taken for granted. This aspect of the learning theory is extensively worked out in computational models formulated within the Universal Grammar tradition:

- cues are defined in extenso and their application is operationalized;
- the exact nature of the interplay between cues and parameters is controlled;
- the order in which parameters are fixed is determined.

Moreover, for a computational model to become operational, problems raised by learnability theory have to be tackled and operationally solved: is parameter setting deterministic or reversable? Is it possible to acquire language without negative evidence, i.e., solely on the basis of positive instances? etc.

In the domain of metrical phonology, computational learning models have recently been formulated: Gupta & Touretzky (1991), Dresher & Kaye (1990), Nyberg (1991). They all approach the problem of how to learn the regularities of stress assignment from the angle of the Universal Grammar theory. In next section we will focus on Dresher & Kaye's computational model of stress acquisition, and we will particularly focus on some plausible alternatives to the nativist approach advocated in that system.

4. NATIVISM VERSUS EMPIRICISM

In previous research (Gillis et al. in press, Daelemans et al. in press) the model developed by Dresher & Kaye (1990, see also Dresher 1992) was critically analyzed along several dimensions. viz. the internal adequacy of the model to explain the acquisition of stress assignment, the adequacy of the notion that parameters distinguish possible and impossible stress systems. These elements will be briefly summarized in this section and we will add a third dimension: at least in the domain under consideration a nativist approach can be replaced by an empiricist approach that makes minimal assumptions concerning a priori knowledge.

- Notwithstanding the impressive innate or a priori knowledge brought to bear on the task in the nativist approach, we found that the model was unable to deal in a satisfactory way with the intricacies of the Dutch stress system, esp. it was shown that the setting of particular parameters is questionable, if not impossible. Moreover, it was shown that the model proposed by Dresher & Kaye is not able to deal with exceptions and language specific rules in a satisfactory way.

- The notion of a set of parameters that describe the possible stress systems can be questioned. Does the set of parameters distinguish possible (existing) from impossible (non-existing) stress systems? Gupta & Touretzky (1991) point out that Dresher & Kaye's set of parameters is not able to describe numerous existing systems (a possibility Dresher & Kaye are well aware of, see e.g., p. 175) This means that the parameter set does not discriminate existing from non-existing patterns. It also implies that more parameters should be added to the set in order to cover the range of possible stress systems. As such, the solution appears to be straightforward and quite easily implemented.

However it can be argued that the inventory of parameters is ultimately based on distributional evidence, i.e., which types of stress patterns have been attested in the languages investigated? And which general characteristics (and, hence, parameters) can be abstracted from them? This leaves the possibility that no single set of parameters will ever be exhaustive since new, as yet unknown, facts may turn up (unless the set of parameters and the possible settings is motivated by theory internal considerations).

This leads to the conclusion that eventually a data-driven or empiricist approach may prove to be more fruitful because it starts from observations of the input data, and works its way up to generalizations. Moreover a similar approach may be supplemented at some point by a priori considerations.

- Is the use of parameters really necessary? Dresher & Kaye (1990:146) argue that "A rich and highly structured theory of UG is otiose if the same results can be achieved by simpler means." What might these alternatives be? One of the alternatives can be described as follows: it appears that stress patterns are sensitive to sequences of syllables and syllable weight (cf. section 2).[2] 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. The table generated in this manner would constitute the grammar of stress for a language.

Dresher & Kaye (1990) argue against such a theory. The argument against a similar approach is that it does not exclude "Crazy nonexistent" (148) and unnatural overall patterns and/or that it allows for "systems with no discernible pattern at all." (148) Metrical theory on the other hand provides a priori constraints on the range of possible stress strings: only certain possibilities are (or better, can be) considered instead of an unstructured range. In other words, parameter settings determine what are possible stress systems: "A compelling argument against such a grammar has to do with the distribution of observed stress systems, as opposed to non-existent ones. The metrical theory (...) allows for a limited number of basic stress patterns." (148)

Thus the claim is that the 'table-look-up' method does not place any restrictions on the hypotheses the learner can entertain. Metrical parameters, on the contrary, determine a priori which stress systems are possible and which are not.

The first part of the claim is obviously correct. Viewed as an enumeration of observed mappings from weight strings to stress strings, the table-look-up method does not a priori exclude any mappings in the way parameter theory does. However it has the advantage that it is restricted to patterns actually observed in the input. As to the second part of the claim - parameters distinguish existing from non-existing patterns

[2] Metrical analyses of Dutch assume four levels of syllable weight: super light (schwa), light (VV), heavy (VC), and super heavy (VCC and VVC). Super lights are never stressed, super heavies are always stressed.

- it is assumed that the parameters and their possible settings exhaust the range of possibilities. However, we already pointed out that this assumption is not warranted by the empirical facts and that ultimately the definition of the parameters is based on distributional phenomena, which is exactly the kind of approach taken in the proposed alternative.

The gist of Dresher & Kaye's (and for that matter, any other nativist approach to language acquisition) is: "Why should stress systems be confined to such a small part of the solution space if learning is based on unrestricted weight-string to stress-string mappings?" (p. 150) From a typological point of view this is a legitimate question indeed: why is the number of observed stress systems in the languages of the world fairly restricted? However, from the point of view of the learner, the question is completely irrelevant since (s)he is confronted with (literally: observes) only one stress system, viz. the system of the language (s)he is exposed to. Hence, if it can be shown that the alternative method proposed can accurately learn the stress system of a language without recourse to an extensive body of a priori knowledge, it is to be preferred as an account of the acquisition of stress assignment (though not necessarily as a theoretically motivated account of typological facts).

As to the learning theory associated with an approach that maps weight strings onto stress strings, viz. keeping a table of weight strings and entering the corresponding stress strings as the relevant data come in, Dresher and Kaye admit that the table will not be very large for any particular language. However they note the following: "... such a learner only appears to learn stress patterns; what it learns is not the pattern itself, but only a part of its extension. It would be unable to assign stress to weight strings not yet encountered. We conclude that such learning theories are empirically inadequate." (Dresher & Kaye 1990:150) In this form the criticism is ad rem. But it assumes that the computational model solely consists of a learning component. It is useful however to distinguish between a performance component and a learning component. The performance component performs an input-output mapping that approximates the target mapping. The structure of these performance mappings is determined by the knowledge accumulated by a learning component. Dresher & Kaye do not seem to take this difference into account (although their own model incorporates a similar distinction between what they call a Learner and an Applier). Moreover, they appear to assume that the performance component applies exact matches to the patterns stored in a table. In other words, it presupposes that there is no mechanism to perform 'best' matches in the absence of an exact match.

The research reported in this paper aims at exploring the potential of learning algorithms that share a data-driven (empiricist) mode of learning instead of the nativist approach. The mechanisms can be seen as instantiations of the learning theory that Dresher & Kaye characterize as "empirically inadequate". We will aim at elucidating the following issues:

- How far can we get in acquiring noise-tolerant generalizations?
- How far can we get without presupposing a priori knowledge: given the fact that even in data driven algorithms knowledge is present in the encodings, we want to strip the encodings from a priori knowledge as far as possible.
- How far can we get when we force the systems not to apply exact matches but to look for the closest match they can find.

If these issues can be treated in a satisfactory way, we believe that the potential of an empiricist method as opposed to the strong nativist approach has been empirically shown, and hence meriting a substantial and thorough exploration in other areas of the linguistic system.

5. Experiments

5.1. Method

In all experiments the leaving-one-out method was used. For this purpose, each item in the dataset in turn is selected 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 expensive method has as its major advantage that it provides the bast possible estimate of the true error rate of a learning algorithm (Weiss & Kulikowski 1991).

5.2. Data and Data Coding

For the purpose of this study a lexicon of Dutch monomorphematic (or morphologically simplex) words was compiled. The lexicon consists of 4868 polysyllabic monomorphemes. It was extracted from the CELEX lexical database which contains 130,778 lemmas and 399,816 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). Only words that could be unambiguously characterized as monomorphemes were selected for our data set. Proper nouns were withdrawn from the dataset. As such our lexicon constitutes a representative sample of the monomorphemes of the language.

The data were encoded according to four coding schemes:

i. The first encoding used the notion of syllable weight: words were encoded as strings of syllables weights, i.e., 1 stands for super light, 2 for light, 3 for heavy, and 4 and 5 for the two super heavy types;

ii. Words were encoded in terms of their rime projections;

iii. No coding categories were added to the phonemic representation, except that only those elements that appeared to be the most information bearing (as determined by the information gain measures) were encoded: viz. the nucleus of the ultimate, penultimate and antepenultimate syllables and the coda of the ultimate syllable;

iv. No encoding was performed, i.e., the complete phonemic representation was used as such.

Of each word only the ultimate, penultimate and antepenultimate syllables were included in the data encoding for convenience sake. The four encodings applied to the word *nirvana* look as follows:

	Syllable		
	Antepenultimate	Penultimate	Ultimate
	nir-	**-va-**	**-na**
Encoding-1	3	2	2
Encoding-2	I r	a -	a -
Encoding-3	I	a	a -
Encoding-4	nIr	va	na

5.3. ALGORITHMS

We have experimented with two similarity-based symbolic learning algorithms: Analogical Modeling (ANA) and Instance-Based Learning (IBL). The two algorithms are supervised (a number of training items is provided). IBL is an incremental algorithm, ANA is a batch learning algorithm. In both algorithms, similarity plays a central role: similar instances have similar categories. Both IBL and ANA make explicit use of similarity-based reasoning. They use a similarity metric to compare items, and use the items most similar to a test item as a basis for making a decision about the category of the latter.

5.3.1. INSTANCE-BASED LEARNING

Instance-Based Learning (Aha et al. 1991) is a framework and methodology for incremental supervised machine learning. The distinguishing feature of IBL is that no explicit abstractions are constructed on the basis of the training examples during the training phase. Instead a selection of the training items themselves is used to classify new inputs. IBL shares with Memory-Based Reasoning (Stanfill & Waltz 1989) and Case-Based Reasoning (Riesbeck & 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 abstraction extracted from this experience (e.g., in the form of rules or decision trees). In linguistics a similar emphasis on analogy to stored examples instead of explicit but inaccessible rules is present in the work of a.o. Derwing & Skousen (1989). As far as algorithms are concerned, IBL finds its inspiration in statistical pattern recognition, especially the rich research tradition on the nearest-neighbour decision rule (see e.g., Devijver & Kittler 1982 for an overview).

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 the 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 the training items to be kept in memory, and the similarity metric used. In these experiments, we "remembered" all training items, and 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).

5.3.2. ANALOGICAL MODELING

Analogical Modeling (Skousen 1989) is another similarity-based framework, meant to provide an alternative to rule-based linguistic descriptions as a model of actual language usage. The main assumption underlying this approach is that many aspects of speaker performance are better accounted for in terms of 'analogy', i.e., the identification of similarities and differences with other forms in the lexicon, than by referring to explicit but inaccessible rules.

The rather vague notion of 'analogy' is given an operational definition in terms of a matching process between an input pattern and a database of stored exemplars. The result of this matching process is a collection of examples called the analogical set. Classification of the input pattern is achieved through selection from this set. The probability that a specific pattern will serve as the analogical example depends on several interrelated factors, such as its frequency, its similarity to the input pattern and the availability of more similar patterns with the same distribution of categories. Analogical Modeling thus shares a number of important characteristics with IBL: in both approaches the main repository of knowledge is a database of stored examples. These examples themselves are used to classify new items, without intermediate abstractions in the form of rules. In order to achieve this, both have to perform an exhaustive search of the database and apply a measure of similarity to retrieve the most relevant items. The main difference between the two approaches concerns the way in which the selection of relevant examples is made: in out information-gain extension of IBL, different weights are attached to each feature in a pattern, thus favouring correspondences between informative features over similarities between less important features. ANA instead relies on the notion of supracontextual heterogeneity to discard irrelevant examples.

5.4. EXPERIMENT 1

In a first experiment we wanted to find out how far we can get in acquiring noise-tolerant generalizations using the data-driven algorithms. Moreover we wanted to measure the influence of the various encodings on the global performance of the algorithms. Given the fact that even in data driven algorithms knowledge is present in the encodings, we stripped the encodings from a priori knowledge as far as possible. It was hypothesized that if a priori knowledge is necessary the results for the first encoding would be superior to the results for the three other encodings.

5.4.1. RESULTS

The rationale behind the encodings used can be described as follows:

i. Encoding-1 uses categories that are thought to be indespensible in metrical theory for describing stress assignment. As such this coding scheme respresents a very strong a priorism. Moreover it is exactly this encoding that Dresher & Kaye (1990: 146) propose in their discussion of the alternatives to the parameter account.[3]

ii. The second encoding consists of rime projections. It coincides with the first one in this sense that it provides only those phonetic or phonemic elements on the basis of which syllable weight is determined.

iii. The third encoding includes those elements filtered through information gain: if a learner is sensitive to the predictive power of distributional facts, a measure of information gain may prove to be a good estimate of the weighted importance the learner assigns to particular features.

iv. The fourth encoding is nothing more than a phonemic representation of the word without any theoretically impregnated manipulations. As such it can be conceived as a very close representation of the bare input to the 'natural' learner.

The first encoding is the one that carries most a prioris, while the fourth encoding is not biased by any a prioris (except for the fact that a word is inputted in a syllabified form). The second and the third encoding take a middle position: the second encoding is manipulated with the assignment of syllable weight in mind, while the third encoding is filtered in an information theoretical fashion.

[3] Dresher & Kaye (1990: 146) note more specifically: "Let us assume that any theory of stress ought to encode at least that much information, so we would not have to consider the infinite number of possible but nonexistent rules which relate stress to other features of the phonetic string." The latter 'encoding' is what we refer to as encoding-4.

A general comparison of the results is shown in Figure 3: the success rates for the two algorithms are plotted out for each encoding scheme. Both algorithms are fairly successfull: for the four coding schemes the success scores are inbetween 80 and 90%, with peak performances of 88.23 for IBL-3 (IBL trained with material encoded using the third coding scheme) and 88.48% for ANA-4.

Insert Figure 3 about here

The results show that for both algorithms the first encoding scheme scores less well than the others, i.e., the theoretically impregnated coding scheme in terms of syllable weights is not as powerful in predicting stress. The scores for the first encoding are significantly lower than the scores for the other encodings ($p < .0001$ in all comparisons).

The other encodings can be said to be largely equivalent in the case of IBL: the differences between the success scores in absolute terms are statistically not significant. In the case of ANA the results for second and the third encoding do not differ significantly, but the fourth does yield better results ($p < .01$).

Taken together these results indicate that an empiricist data-driven approach to the problem of stress assignment is feasible. Moreover, the less the input representations are biased by theoretically motivated considerations, the better the results turn out to be. Especially the comparison of the results of the first and the second encodings are highly revealing: syllable weight (encoding-1) is determined on the basis of rime-projections.(encoding-2). The latter yields a significantly better performance though, which implies that in abstracting syllable weight necessary information to solve the task at hand is lost.

The question now crops up how well these results are in comparison to a computational model that incorporates a nativist theory, such as Dresher & Kaye's model. We did not implement the latter acquisition model, however, a good basis for comparison is the following: a state of the art metrical analysis was performed for each entry in the lexicon. It was determined if a word was Regular in this respect or required an idiosyncratic lexical marking. Since our analysis of the model (Gillis et al. in press) showed that even for Regular words the Dresher & Kaye model might run into problems (unless a brute force learner is invoked that checks all possible parameter settings), the number of Regular words may turn out to be a fairly good estimate of the success rate of the model. Given this assumption, the 'nativist' model would score less well then the two 'empiricist' models: the number of Regular words amounts to 3916 (80.44%) and the number of words requiring idiosyncratic lexical marking amounts to 952 (19.56 %). The least we can conclude from these figures is that the 'nativist' model does not score significantly better than the 'empiricist' models.

5.5. EXPERIMENT 2

The results of the first experiment can be interpreted as denoting that the more elaborate the encoding the less ambiguous the input patterns become. As a consequence the patterns are stored in memory with a single target category associated with them and thus no generalization whatsoever is performed. In Table 1 the number of elements specified in each encoding is specified together with the number of different patterns this results in and the number of those patterns that are unambiguous. It can be readily seen that encoding-1 contains less elements than encoding-3, which in its turn contains less elements than encoding-2, and encoding-4 is the most detailed. If we take the fourth encoding, it appears that only one single pattern is ambiguous. Thus if the system's performance would be solely based on memory look-up, a success-rate of 99.98% would be attained, which is certainly not the case. The algorithms do generalize in some sense.

Table 1: Overview of the fine-grainedness and associated ambiguity of the four encodings

Encoding	Number of Elements Encoded	Number of Patterns	Number of Unambiguous Patterns
Encoding-1	3	87	45
Encoding-2	6	2724	2607
Encoding-3	4	2006	1846
Encoding-4	9	4818	4817

But again it is striking that for analogical modeling, the more elements are included in the encoding the better the performance becomes (in absolute terms). IBL appears to be less sensitive to this: the third encoding, in which only four elements of the input word are withheld, scores better than the second and the fourth encodings.

In our second experiment we wanted to get rid of the idea that - as Dresher and Kaye implied - there is no genuine generalization in the performance of the type of algorithm exemplified by analogical modeling and instance based learning. For this purpose the performance of the algorithms was manipulated in such a way that in the testing phase no exact matches between the exemplars in memory and the test-item were allowed. This condition is called the 'forced generalization condition' in contradistinction with the 'regular generalization condition' used in the first experiment.

5.5.1. RESULTS

The results of the simulations are displayed separately for IBL and ANA in Figure 4 and Figure 5. Well in agreement with what could be expected, the results for the 'forced generalization condition' are less well than in the 'regular generalization condition'. The differences are significant at the 1% level or below. The noteable exception is the phonemic encoding (encoding-4): the difference between the results in the two conditions is not significant for the two algorithms. In the case of IBL the 'forced generalization' result is even better in absolute terms than the result in the 'regular generalization condition'.

Insert Figure 4 and Figure 5 about here

The results of this experiment indicate that even if the possibility of exact matches in the testing phase is eliminated, the algorithms are able to form generalizations. That is, contrary to the claim of Dresher & Kaye, an alternative to the nativist approach can deal with input patterns that cannot be retrieved from memory immediately. Moreover if we apply the same test as in the previous experiment, the results of this forced generalization experiment provides evidence that the performance of these algorithms is still well above the performance of the model advocated by Dresher & Kaye.

6. CONCLUSION

Linguists and psycholinguists tend to agree that part of the knowledge humans bring to the task of language acquisition is innate. Chomskyan linguists go a long way along this path, arguing that Universal Grammar, the fundamental concepts and principles operative in the world's languages moulds the acquisition process. In other words the principles and parameters approach to language acquisition holds that the main structural aspects of the linguistic system are acquired through a process of parameter setting. These parameters are innately given. From an analysis of a computational model of the acquisition of metrical phenomena it was concluded that if the parameters are innately given, the cues for setting the parameters must also be part of an innate program.

In two experiments the viability of an empiricist approach was investigated in the domain of metrical phonology. It appeared that Instance-Based Learning and Analogical Modeling, two algorithms that rely heavily on similarity between an input items and a database of training material, are able to learn the stress patterns of Dutch. Moreover, the algorithms reached superior results when trained with data that were not coded in a theory impregnated way (in contrast with the predictions of the 'nativist' approach).

A second claim in the literature was experimentally falsified, viz. that the kind of approach applied in the experiments would run into trouble when no exhaustive enumeration of the possible stress patterns of a language is available. In our second experiment we showed that even when exact matches with memorized items were eliminated from the algorithms' performance component, the success rate of the algorithms was highly comparable to the normal condition in which exact matches were allowed, provided that a phonemic representation of the words was provided as input. The latter is the representation that comes closest to the input to the 'natural' natural language learner.

The results of these experiments should not be taken as cast-iron proof against nativist approaches to language acquisition. Instead, they show the potential of even fairly simple data-driven algorithms which have been neglected in the language acquisition literature since the Chomskyan revolution. Now that computational models can be used as powerful tools for simulating vast areas of theoretical coverage using large sets of training and test material, these simulations may be a good starting point for exploring the alternatives of the nativist approaches.

REFERENCES

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

Chomsky, N. 1981. Principles and Parameters in Syntactic Theory. In N. Hornstein & D. Lightfoot (Eds.) *Explanations in Linguistics*. London: Longman.

Daelemans, W. & van den Bosch, A. 1992. Generalization Performance of Backpropagation Learning on a Syllabification Task. In: M.F.J. Drossaers and A. Nijholt (eds.) *Connectionism And Natural Language Processing. Proceedings Third Twente Workshop On Language Technology*, pp. 27-38.

Daelemans, W. Gillis, S., Durieux, G. & van den Bosch, A. in press. Learnability and markedness in data-driven acquisition of stress. ITK

Derwing, B. L. & Skousen, R. 1989. Real Time Morphology: Symbolic Rules or Analogical Networks. *Berkeley Linguistic Society* 15: 48-62.

Devijver, P.A. & Kittler, J. 1982. *Pattern Recognition. A Statistical Approach*. London: Prentice-Hall.

Dresher, E. 1992. A learning model for a parametric theory of phonology. In Levine, R. (Ed.) *Formal grammar: Theory and implementation.* Oxford: Oxford University Press.

Dresher, E., and Kaye, J. 1990. A Computational Learning Model of Metrical Phonology. *Cognition* 34: 137-195.

Gillis, S., Daelemans, W., Durieux, G. & van den Bosch, A. in press. Learnability and markedness: Dutch stress assignment. Proceedings of the 15th Annual Meeting of the Cognitive Science Society.

Gupta, P., and Touretzky, D. 1991. Connectionist Models and Linguistic Theory. Unpublished Ms.

Hyams, N. 1986. *Language acquisition and the theory of parameters*. Dordrecht: Reidel.

Nyberg, E. 1992. A Non-deterministic, Success-driven Model of Parameter Setting in Language Acquisition. Ph.D.Diss., Dept. of Philosophy, Carnegie Mellon University.

Pinker, S. 1984. *Language learnability and language development.* Cambridge: Harvard University Press.

Pinker, S. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge: MIT Press.

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

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

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

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

Trommelen, M. & Zonneveld, W. 1989. *Klemtoon En Metrische Fonologie*. Muiderberg: Coutinho.

Trommelen, M. & Zonneveld, W. 1990. *Stress In English And Dutch: A Comparison*. Dutch Working Papers in English Language and Linguistics 17.

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

Figure 3: Comparison of the Global Results for IBL and ANA

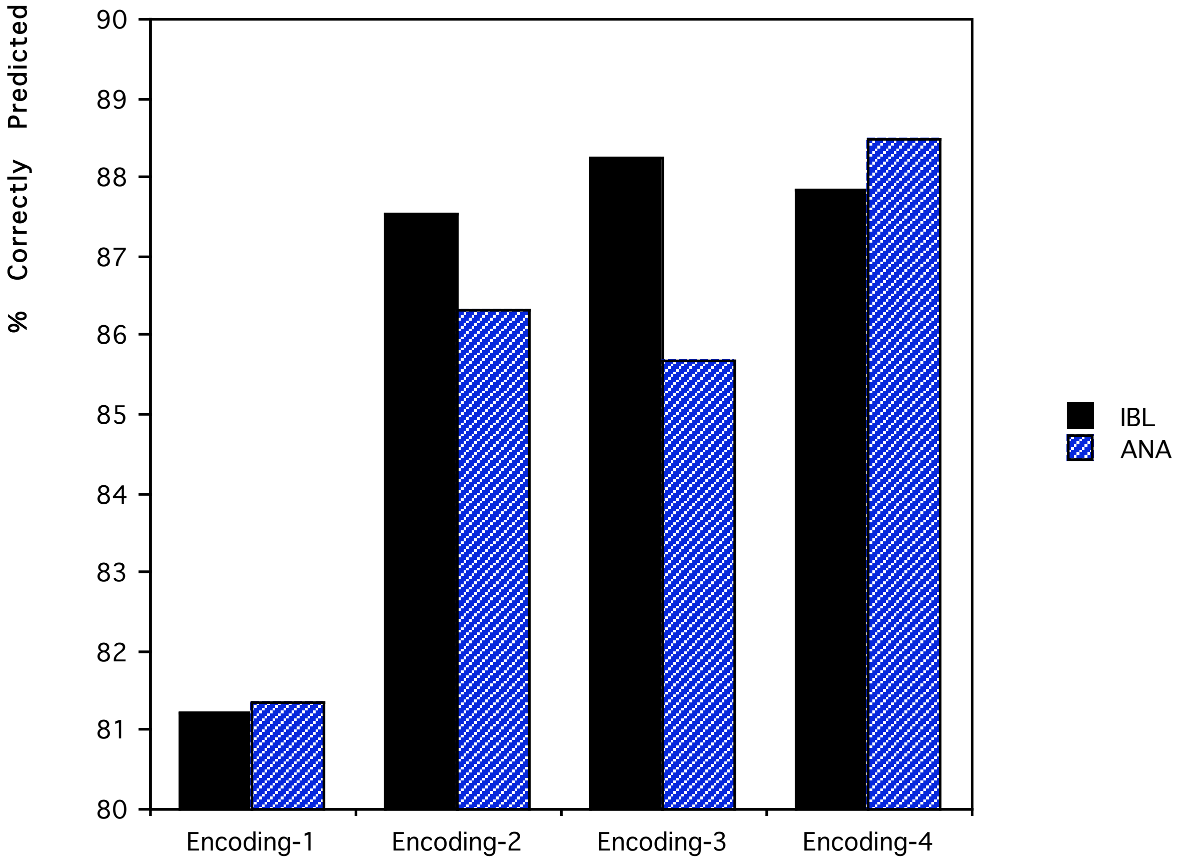


Figure 4

Results of the 'regular generalization' and 'forced generalization' experiments for Analogical Modeling

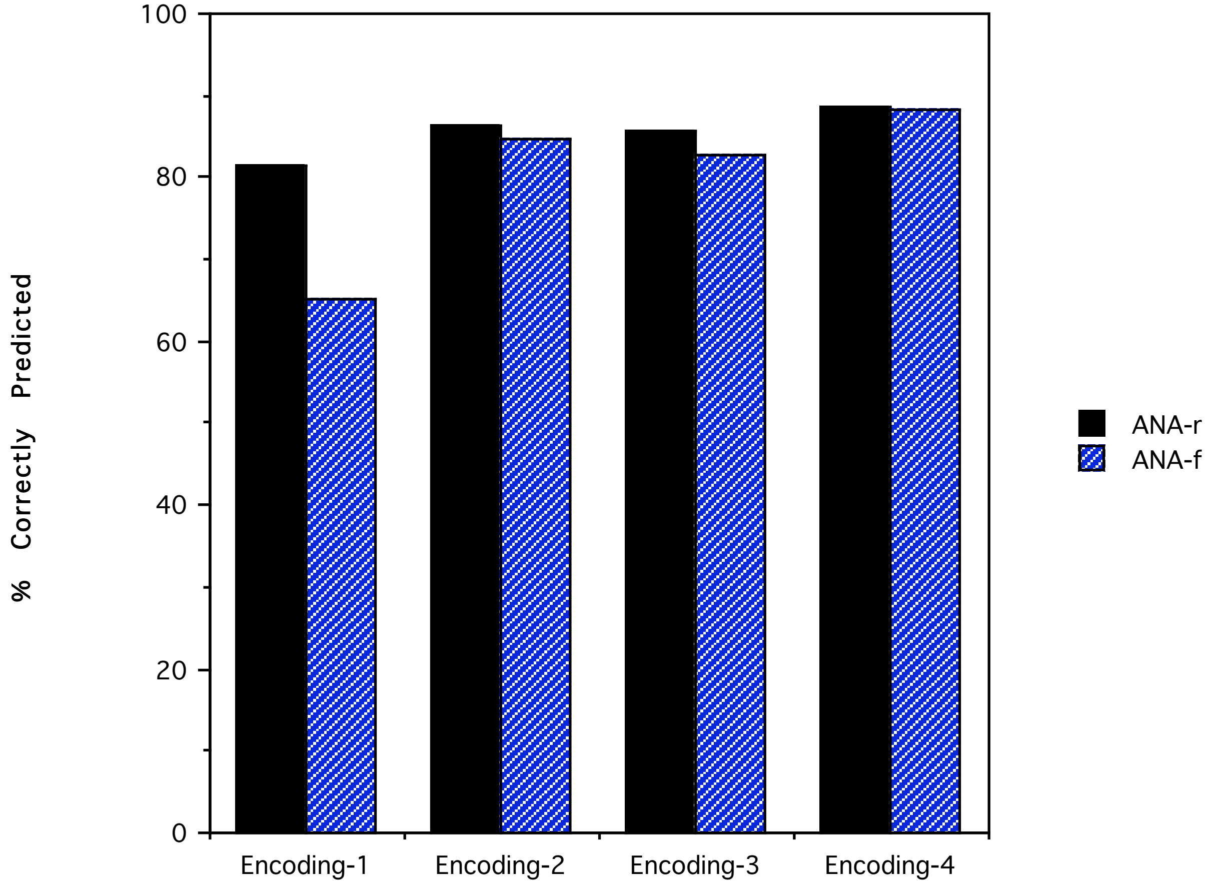
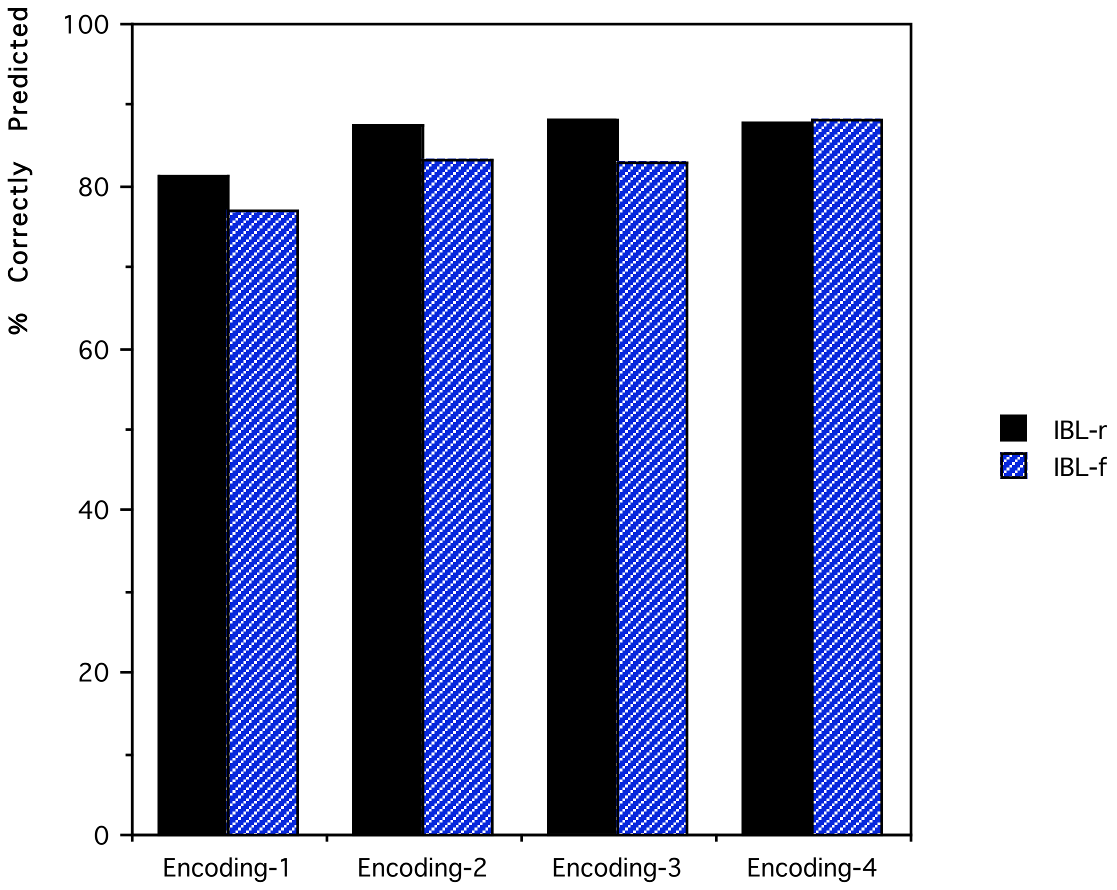


Figure 5

Results of the 'regular generalization' and 'forced generalization' experiments for Instance Based Learning


% Correctly Predicted
Encoding-1
Encoding-2
Encoding-3
Encoding-4
IBL-r
IBL-f