On the Arbitrariness of Lexical Categories

Gert Durieux, Walter Daelemans and Steven Gillis

University of Antwerp

Abstract

In this paper, we look at lexical categories and their predictability from a machine learning perspective. Starting from linguistic intuitions about predictability in three different domains, we show how standard techniques for analyzing classification tasks arive at a similar predictability scale. In the second part of the paper, we carry out machine learning experiments covering these domains and relate learnability results to the previous analysis. The IB1-IG classifier is found to be capable of learning in all three domains, although with varying degrees of success.

1 Introduction

Traditionally, lexical categories are interpreted as being either arbitrary, predictable or partly predictable. Arbitrary categories (e.g. the possible syntactic categories of a word) have to be rote-learned, and are stored in the mental lexicon. Predictable categories (e.g. the proper syllabification of words) can be computed when needed and do not have to be stored. Instead, the required computation can be summarized by stating a number of rules, which map one representation into another. Partly predictable categories, finally, can be computed for the majority, or at least a sufficiently large number of lexical items, but need lexical marking for those cases that the rules do not apply for. The past tense of verbs in languages such as English and Dutch would be an example of this type of category. For these kinds of categories, dual route models are usually proposed. These models delegate exceptional items to the lexicon, and handle the regular cases by a rule component. In recent versions of this approach, a limited form of generalization can also take place in the lexical component.

None of the distinctions made above, however, is absolute: even for predictable categories, some exceptional items may exist, and, conversely, for arbitrary categories, minor generalizations can often be made. Rather, the predictable/arbitrary distinction forms a continuum, along which lexical categories can be placed. A second point to take note of is that claims about the predictability of a lexical category are always made with respect to particular representations: changing the representation often has considerable impact on the degree of predictability of the category in question. A good example is provided by Trommelen's (1983) analysis of diminutive formation in Dutch: by giving an account of this process in terms of the syllable structure of Dutch words, a number of facts can be explained which remain puzzling when the analysis is made only in terms of segmental features (such as vowel length and sonorance) and stress.

In this paper, we explore the predictable/arbitrary dimension from a data-analysis and machine learning perspective. Our hypotheses are (i) that the de-

gree to which a lexical category is predictable can be shown quantitatively by standard data-analysis and machine learning techniques, and (ii) that Memory-Based Learning succeeds in successfully learning lexical categories, irrespective of their place on the predictable/arbitrary continuum. The Memory-Based Learning paradigm (MBL) is based on the assumption that records of past experience, together with a suitably defined measure of similarity, are sufficient to produce intelligent behaviour. Its distinguishing characteristic is that no intermediate levels of abstraction (e.g. rules) are created during learning, but that the stored examples themselves are used directly to categorize unseen examples. Thus, the main attraction of MBL in the context of lexical categories, is that a single architecture could potentially account for phenomena for which either rules, rote-learning or dual route models have been proposed, and could do so using a uniform representation for both regular and exceptional items. Another advantage could be that the problem of constructing appropriate rule sets, or blocking rule application in dual route models is sidestepped.

To explore these hypotheses, we conducted a series of machine learning experiments involving three different lexical categories for Dutch nouns, viz. diminutive, stress and gender. These categories are meant to represent different points on the predictability scale: diminutive formation is assumed to be completely predictable. Stress assignment is slightly more complex, and needs various degrees of lexical marking in order to acount for a sizeable portion of the lexicon (about 20%, depending on the analysis); the degree of marking varies from the addition of a single exception feature to full lexical listing. Gender assignment, finally, is taken to be essentially arbitrary, a few minor generalizations notwithstanding. Each of these processes was cast as a classification task, for which data sets were constructed on the basis of the CELEX lexical database. These data sets were first analyzed using entropy-based measures and machine-learning techniques, in order to verify our first hypothesis, and subsequently used in machine learning experiments with a variant of MBL, to test our second hypothesis.

The remainder of this paper will be structured as follows: in section 2, we will briefly present the linguistic domains chosen, and describe the construction of the data sets. In the next section, we will introduce a number of information theoretic measures and discuss their application to the data sets constructed. Then, in section 4 we will review a series of machine learning experiments and discuss the results in the light of the previous analyses. Finally, we will discuss our findings w.r.t. the hypotheses above.

2 Description of the application domains

In order to test our hypotheses, we selected three different lexical categories, diminutives, stress and gender, in a single language, Dutch. These categories represent different points on the predictability scale, ranging from completely predictable (diminutive) to essentially arbitrary (gender), with a middle position occupied by stress. In the next paragraphs, we will briefly describe each of these categories. After that, we will discuss the creation of data sets for each domain.

2.1 Diminutives, stress and gender in Dutch

The first application domain we selected is diminutive formation. Diminutive formation is a productive process in Dutch, applying essentially to all nouns for which it makes semantic sense. Diminutives are formed by attaching an alternant of the diminutive suffix to the noun stem. In Standard Dutch, there are five such alternants: *-tje*, *-etje*, *-pje*, *-kje* and *-je*. Of these alternants, the first one is generally considered basic, since it is the form encountered with nouns ending in a vowel, and the surface /t/ would otherwise be unmotivated. Three of the other alternants are generated by rules of allomorphy, i.e. rules which effect phonological changes at specific morpheme boundaries only, and are not otherwise encountered in the phonology of the language: instances of such rules for diminutive formation are the schwa-insertion rule which generates the *-etje* alternant, and the assimilation rules which produce either *-pje* or *-kje*. The remaining alternant *-je* is generated by a phonological rule of /t/-deletion. In earlier generative accounts (see Trommelen (1983, chapter 2) for an overview, and Gussenhoven and Jacobs (1998, chapter 7) for a reformulation and further references), these rules were stated mainly in terms of segmental features such as sonorance and vowel length, with occasional reference to suprasegmental features (stress) or morphological structure. Trommelen (1983) has shown that much of the phonological detail can be done away with when structural properties of the final rime are considered, and that reference to stress becomes completely superfluous. More important for our purposes than the details of each analysis, is the fact that diminutive formation is a local process, which is completely determined by the final syllable of nouns, and to which very few exceptions exist.

The second application domain is main stress assignment. Stress in Dutch is rather complicated, being neither fixed as in e.g. French, nor completely free as in e.g. Russian. Although the facts may seem confusing at first sight, they appear to be governed by a number of major and minor generalizations. For underived words, the main generalizations are stated as follows (Kager 1989, 227):

1. Main stress falls within a three syllable window, counting from the righthand word-edge.

2. If the word contains a syllable with an underlying schwa, preceded by a consonant, then main stress falls on the immediately preceding syllable.

3. Main stress never falls on the antepenultimate syllable if the penult is closed or contains a diphthong.

In addition to these major generalizations, for which only a handful of exceptions can be found, a number of minor generalizations have been made. They are related to the CV structure of the final syllable, and only describe strong tendencies; violations of these rules are much more common than exceptions to the major generalizations above:

1. Words ending in a superheavy syllable have final stress.
2. Words ending in a diphthong have final stress.
3. Words ending in a closed syllable have antepenultimate stress, or penultimate stress if they are disyllabic.
4. Words ending in a final open syllable have penultimate stress.

Within the framework of metrical phonology, stress is seen as a relational property, encoding prominence relations among metrical constituents. These constituents are hierarchically organized, building on the syllable (or mora) structure of words. At a higher level, syllables (or moras) are grouped into one or more lexical feet, which together form words. A metrical analysis of a stress system proceeds by specifying the instructions for building up this structure. For Dutch, a number of proposals have been made, which are surveyed in Kager (1989) or Booij (1995). Although the details of analysis vary, in most versions about 80% of words are regular (i.e. the right stresses are generated by normal rule application), whereas the other 20% are handled by lexical marking (see Neijt and Van Heuven (1992) for a quantitative comparison of different proposals). For compounds and derived words, additional complications arise: compound stress is morphologically governed, and main stress falls on the first member of the compound. For affixed words, the stress behaviour depends on the type of affix: level-2 affixes can be either stress-neutral, stress-bearing or stress-attracting. Stress-neutral affixes leave the stress pattern of the underived word untouched, stress-bearing affixes always bear main stress themselves, and stress-attracting affixes lead to shifts in the stress pattern of the underived word. Comparing stress to the previous domain, then, it appears that stress is a non-local phenomenon, which is only partly predictable from the phonological structure of words.

The last application domain is gender. Historically, Dutch had a three-gender system, distinguishing the traditional categories of *masculine*, *feminine* and *neuter*. In its current state, the system has largely completed the transition towards a two-gender system, losing the distinction between masculine and feminine along the way. Remnants of the three gender system are still observed with pronominal anaphora. In the Netherlands the masculine/feminine distinction is preserved when the antecedents denote persons or, to a lesser degree, animals, and coincides roughly with the male/female distinction. In Flanders, the opposition is observed more strictly, and holds for non-animate antecedents as well. Although the gender of nouns is usually clear from agreement phenomena, gender assignment rules have proven difficult to formulate. Haeseryn et al. (1997, chapter 3.3) list a number of rules based on morphological and semantic criteria, but note that these rules cover only a small part of the noun lexicon, and that exceptions or regional variations abound. Psycholinguistic research (Deutsch and Wijnen 1985; van Berkum 1996) tends to confirm the view that gender is simply stored as part of the lexical information associated with nouns. This has led some researchers to doubt the possibility of solving the gender assignment problem for Dutch: "The relationship between article and noun in Dutch is, except for a few exceptions, more

or less arbitrary: the form the article takes is not systematically determined by any phonological, morphosyntactic, semantic, or conceptual features of the noun." (Deutsch and Wijnen 1985).

2.2 Construction of data sets

In order to test our hypotheses, each of the lexical processes described above was recast as a classification task. This makes them amenable to standard techniques of comparing classification tasks, and will allow to use the same data directly to run experiments with an MBL classifier. Input to such a learning algorithm consists of a number of observations (or instances), which take the form of feature vectors. Feature vectors are ordered sequences of atrribute/value pairs, of which a single one is designated as the target or class attribute. The remaining attributes are called predictor attributes. The task of the classifier is to construct a mapping from predictor to target attributes during an intial learning phase. During a subsequent test phase, the constructed mapping is applied to unseen observations.

Two basic data sets were selected from the CELEX lexical database: the first contains 5000 monomorphemic nouns, varying in length between one and four syllables. For the second data set, 15000 noun lemmas were selected, comprising both monomorphemic and complex (i.e. derivations and compounds) nouns. Rather than use task-specific attributes for each domain, we decided to use a single encoding for all tasks, providing only a minimally required level of lexical information. From CELEX, the phonological transcription of each noun in the data set was extracted. This transcription was then syllabified, and each syllable was further divided into onset, nucleus and coda, yielding a total of twelve predictor attributes per observation. For shorter words, or syllables without onset or coda, unoccupied positions were padded with null values. Data were right-aligned, meaning that all final syllables were lined up.

The data sets were then provided with target attributes for each of the problem domains: for diminutives, the five alternants of the diminutive suffix are the obvious choice, leading to five classes: *-tje*, *-etje*, *-pje*, *-kje* and *-je*. For stress, four target classes were created, *s-1* to *s-4*, denoting stress on the first, second, third or fourth syllable from the right. Gender assignment was split into two classification tasks, one for the two-way gender distinction and one for the traditional three-way gender distinction. For the former, which will from now on be referred to as de/het, there are three classes, named after the form of the definite article: *de* for non-neuter, *het* for neuter and *de/het* for nouns where both genders occur. For the three-way gender distinction, the following classes were used: *m.* for masculine, *v.* for feminine, *o.* for neuter, *v. (m.)* for feminine nouns which also occur as masculine and, finally, *m.-v.* for those nouns were gender is dependent on the sex of the referent (compare the word 'nurse' in English).

3 An information-theoretic analysis

Our first hypothesis was that predictability (or conversely, arbitrariness) can be shown quantitatively, using standard measures for comparing classification tasks. In this section we will analyse the data sets for monomorphemes. The presentation will be incremental: first, we will look at the respective class attributes, then we will consider the predictor attributes, and finally, we will consider the data sets from a more global perspective, focussing on the structure of the application domain.

3.1 Class attributes

In constructing the data sets, care has been taken to make them as similar to each other as possible. Indeed, the number of instances has been kept constant, as has the number and type of predictor attributes. Of course, variation exists both in the number and distributions of the target attribute, which was dictated by the problem domain. Thus, a first step in the analysis, has to be to assess the impact of the target classes and their respective distributions within the data.

Comparing just the number of classes, there are two tasks, diminutive and gender, with five possible classes, stress has four classes and de/het has three. Of course, this is a very crude measure, since it does not take the class distributions into account: classification tasks involving five classes, where one class accounts for 90% of the observations, are not necessarily more difficult than problems with three classes, where all classes are equally likely. For this reason, the information-theoretic measure entropy is widely used in assessing the difficulty of classification problems. Entropy is essentially a measure for the randomness of a random variable, and is given by

$$H(C) = -\sum_{i}^{q} P(C_i) log_2 P(C_i) \quad (1)$$

where $H(C)$ is the class entropy, and $P(C_i)$ the probability that class i occurs. Thus $H(C)$ is maximal when each class is equally likely, and ranges from 0 for a single class to $log_2 q$ for q classes. A useful way to look at entropy is to take $2^{H(C)}$ as the effective number of classes. Applying this measure to the data sets yields the figures in table 1:

	q	$H(C)$	$2^{H(C)}$
diminutive	5	1.52	2.86
stress	4	1.09	2.13
gender	5	1.77	3.41
de/het	3	0.70	1.63

Table 1: Analysis of class attributes

Taking into account class distributions, gender prediction appears to be the hardest task, having both the largest observed and effective number of classes and the highest class entropy. Diminutive formation, which has the same number of classes, but slightly lower values for class entropy and effective number of classes, comes second, followed by stress. De/het, finally, has the lowest values for all three measures, which makes it, intrinsically, the easiest task of all.

3.2 Predictor attributes

Since the classifier is informed by the predictor attributes, it is also important to assess the degree to which these predictor attributes are helpful in determining the class attribute. A common measure to determine this for each predictor variable is to use mutual information (Henery 1994). Mutual information measures the amount of randomness that is removed from the class variable by knowing the value of a predictor variable. It is defined by

$$M(C,X) = \sum_{ij} p_{ij} log_2 \frac{p_{ij}}{P(C_i)P(X_j)} \tag{2}$$

where p_{ij} denotes the joint probability of observing the jth value of attribute X and class C_i, and $P(C_i)$ and $P(X_j)$ the marginal probabilities of observing the ith class and jth value respectively. Mutual information is bounded by $0 \leq M(C,X) \leq min(H(C), H(X))$ where 0 is the value when the class and attribute are completely independent, and the maximal value denotes that the class is completely predictable once the attribute value is known. Under the assumption that all attributes are independent, the total amount of useful information is obtained by summing the mutual information for all of the attributes. Division by the number of attributes r yields the average mutual information:

$$\overline{M}(C,X) = r^{-1} \sum_{i} M(C,X_i) \tag{3}$$

The average mutual information can be used to derive an additional measure: if we take the class entropy as the amount of information which has to be supplied, and view the predictor attributes as contributing individual bits of information, then dividing the class entropy by the average mutual information yields the equivalent number of attributes, $EN.attr$.

$$EN.attr = \frac{H(C)}{\overline{M}(C,X)} \tag{4}$$

This measure indicates how many attributes with the average mutual information would be necessary for classification. If this number exceeds the actual number of attributes, this is an indication that the current set of attributes is insufficient to solve the classification task. It is only an indication, however, since the assumption of independence between attributes is likely to be unrealistic in most cases, and the

combined mutual information of attributes may well exceed the amount obtained by summing the mutual information for each of the attributes.

A final measure we will use for the analysis of the data sets is the noise-to-signal ratio of the attributes. If we take the average mutual information as the amount of useful information supplied, and the difference between the average attribute entropy and the former quantity as the amount of non-useful information contained in the attributes, then the following measures gives the amount of noise (or irrelevant information):

$$NS.ratio = \frac{\overline{H}(X) - \overline{M}(C,X)}{\overline{M}(C,X)} \tag{5}$$

Large values of the noise-to-signal ratio indicate that the chosen attributes contain much irrelevant information, and that such data sets could be condensed without loss in classification accuracy.

Applying the measures introduced above to each of the data sets yields the figures in table 2:

	$\overline{M}(C,X)$	*EN.attr*	*NS.ratio*
diminutive	0.236	6.41	7.06
stress	0.235	4.64	7.08
gender	0.043	41.14	43.23
de/het	0.011	62.41	168.43

Table 2: Analysis of predictor attributes

From this table, a clear dichotomy appears between diminutive and stress on the one hand, and gender and de/het on the other: diminutive and stress have fairly comparable values for both average mutual information and noise-to-signal ratio. The expected number of attributes for both problem domains falls well within the actual number of attributes supplied. Taken together, these figures indicate that the chosen attributes are useful for the classification task at hand, and that the data sets contain relatively little noise. For both gender and de/het, the average mutual information values are considerably smaller, with a corresponding increase in noise-to-signal ratio (recall that the attribute entropy is the same for all data sets). This implies that the encoding used contains little useful and much irrelevant information. The equivalent number of attributes is much higher than the actual number supplied, which indicates that additional information is needed to solve the classification task.

In the light of the linguistic description of the problem domains, these findings come as no surprise: not only are both diminutive formation and stress assignment located on the higher part of the predictability scale, their analysis is cast entirely in terms of phonological properties of the lexical items involved. Although there is no complete overlap between the attributes selected for the classification tasks, and those referred to in the linguistic analysis (e.g. structural information is highly simplified and non-hierarchical in our feature encoding), the base material on which

these analyses build is in place. For gender and de/het the reverse is true. Not only is gender assignment taken to be essentially arbitrary, the proposed rules of thumb do not refer exclusively (if at all) to phonological properties of the relevant lexical items.

Considering tables 1 and 2 together, it appears that our first predictability ranking is in need of qualification. De/het, which was the easiest problem judging from the class entropy, is also the problem for which the predictor attributes contain the least amount of useful information. Gender, which was the most difficult one, also suffers badly from insufficient and irrelevant information. Therefore, at this point of the analysis, they should both be located at the lower end of the predictability ranking. For diminutives and stress, the situation is different. Based on the figures obtained for their class attributes, they occupied a middle ground between gender and de/het. Analysis of the predictor attributes has shown both to have fairly informative attributes, with relatively little noise. Therefore, they should prove easier to predict than both other categories.

3.3 A structural view

So far, we have looked at the class distributions for the different problem domains, and at the information contained in the attributes. What we have not considered yet, is the overall structure of the problem domains. The feature vectors used for the encoding of the observations define an n-dimensional hyperspace, where each observation corresponds to a single point. From a geometrical point of view, the distribution of points in this space can give an indication of the complexity of the domain. In the simplest case, all observations belonging to a single class would be located within the same region, and boundaries between classes could be defined easily. For these kinds of problems, even simple classifiers do well. Complexity increases when there are several regions containing instances of the same class. Such concepts are called disjunctive or polymorphous, and a high degree of polymorphism considerably complicates classification tasks.

Several quantitative measures can be used to show the degree of polymorphism: the number of clusters (i.e. groups of nearest neighbours belonging to the same class), the number of disjunct clusters per class or the number of prototypes per class (Aha 1992). In this paper, we used the learning component of FAMBL (van den Bosch 1998) to analyse our data sets. FAMBL is an instance-merging variant of MBL, which is centered around the notion of instance families. An instance family is a cluster of nearest neighbours, belonging to the same class. In FAMBL, these families are constructed during the learning phase, and subsequently merged into hyperrectangles, which are then used for classification. For our purposes, the number of families created, and the median number of instances merged per family provide some insight into the overall disjunctivity of the data. In table 3, the results for our data sets are summarized:

	#families	*avg.clustersize*
diminutive	41	21
stress	137	7
gender	309	3
de/het	370	5

Table 3: Geometrical analysis

From this table, it can be seen that diminutive is the domain with the lowest disjunctivity: there are 41 families, with a median size of 21. The numbers for stress are higher: here the number of families created is more than three times as high, with a corresponding decrease in the median family size. However, these figures are still well below those for gender and de/het: here the instances are scattered across a multitude of families, with a low median number of merged instances. The implications for the predictability of these domains largely agree with those found during the analysis of the predictor attributes: gender and de/het are the hardest problems, stress should prove easier to predict, and diminutive is the most predictable category.

Our final ranking thus appears to be in perfect agreement with the remarks made in section 2.1. Since it was derived independently, using techniques from data analysis and machine learning, we consider this as evidence in support of our first hypothesis.

4 Predicting lexical cateories

The second hypothesis was that MBL would be able to learn lexical categories successfully, irrespective of their place on the predictability scale. Underlying this hypothesis is the fact that a PAC-learning analysis has shown a simple variant of MBL to be capable of learning a large class of concepts, including highly disjunctive ones (Aha 1990). Daelemans, van den Bosch and Zavrel (1999) have further shown that MBL algorithms which keep all training instances in memory compare favourably to instance-editing or instance-averaging approaches, and outperform decision tree learning methods on a number of linguistic benchmark problems.

4.1 Evaluation of classifiers

Before we turn to the experiments, a few words must be said about evaluation of classifiers. The most commonly used measure for the performance of classifiers is *accuracy*, i.e. the ratio of correct classifications to the total number of predictions made. Accuracy is usually measured on a separate test set, which is disjoint from the training set used during the learning phase. Additionally, to minimize the impact of sample selection, some form of cross-validation is used. Although accuracy is a simple and attractive measure, it has a number of drawbacks which make it less suitable for cross-classifier or cross-domain comparisons. The first drawback is that not all classifiers return the same kinds of answers: some classifiers

return a single class, whereas others return a probability distribution over some or all classes. In the latter case, the answer first has to be interpreted as wrong or correct before accuracy can be calculated. A second drawback is that prior class probabilities are not taken into account: for two class-problems with respective distributions of (0.5 0.5) and (0.9 0.1), an accuracy of 70% on the first one would be a relatively good score, whereas the same accuracy on the second problem is inferior to what would be obtained by always guessing the more frequent class; the difference, however, is not brought out by the accuracy scores, which are the same. A similar problem arises when domains with different numbers of classes are compared.

Fortunately, there are information based measures which overcome these defects. The nexts paragraphs will describe one such measure, which is due to Kononenko and Bratko (1991). The starting point for the derivation of their measure is the observation that a classifier's answer can be considered useful, if it makes the true class of an instance more probable than the a priori probability of that class. Conversely, an answer is considered misleading when it makes the true class less probable than its a priori distribution. Of course, a correct answer is also useful, and a wrong answer is also misleading. To measure a classifier's performance, useful answers should be rewarded, and misleading answers should be penalized.

The amount of credit/penalty which should be assigned is related to the amount of information that the classifier provides: if the anwer is useful, its *information score* is defined by:

$$I_{answer} = -logP(C) + logP'(C) \tag{6}$$

where $P(C)$ is the prior probability of class C, and $P'(C)$ is the posterior probability returned by the classifier. If the answer is misleading, the penalty is defined by:

$$I_{answer} = -\left(-log(1 - P(C)) + log(1 - P'(C))\right) \tag{7}$$

The average information score I_a is obtained by summing the information scores of all answers, and dividing by the total number of answers given. To relate the average information score to the class entropy of the problem domain, the relative information I_r score can be defined by:

$$I_r = I_a / H(C) * 100\% \tag{8}$$

This last measure is interesting, since it relates the classifier's performance directly to the uncertainty associated with the task. In other words, I_r proovides an indication of the degree to which the classifier has succeeded in solving the classification task.

4.2 Experiment 1

To test our second hypothesis, the data described in section 2.2 were used in machine learning experiments with the IB1-IG classifier. IB1-IG (Daelemans and van

den Bosch 1992; Daelemans, van den Bosch and Weijters 1997) is an MBL algorithm, which keeps all training items in memory. During the test phase, unseen instances are classified by comparing them to all stored instances and retrieving the k most similar items (nearest neighbours). The most frequent class among these k neighbours is then predicted as the class of the unseen instance. The similarity metric used is a weighted overlap metric, with mutual information used as weights.

The test regime used was ten-fold cross-validation, and results were averaged over ten partitions. A summary of the results is displayed in table 4:

	$H(C)$	*accuracy*	I_a	I_r
diminutive	1.52	98.50%	1.46	96%
stress	1.09	92.94%	0.88	80%
gender	1.77	40.72%	0.32	17%
de/het	0.70	73.32%	-0.05	-6%

Table 4: Results for Experiment 1

Looking at the results for the experiments with monomorphemes, a number of observations can be made: the first is that diminutives are learned almost perfectly, with a success rate of 98.50%. This is a clear indication that rule-like behaviour can be produced by an MBL system, and drives home the point we made in the introduction about sidestepping the problem of rule construction. Stress too is learned quite well, with an accuracy of 92.94%. Considering the fact that about 80% of the data was taken to be regular, this shows that for partly predictable categories, an MBL approach can get a sizeable portion of the irregular items right as well. These findings are generally confirmed by looking at the relative information scores: for diminutives, this score is 96%, meaning that the classifier has solved the problem almost completely. For stress, the relative information score is lower, but still respectable.

On the other hand, predictive accuracy for both gender and de/het is low; although 73.32% for de/het may seem reasonable, this figure is misleading, as can be seen from the relative information score, which is negative. This means that even a completely uninformed classifier would do better by always guessing the most frequent category. What we seem to have thus, is a sharp contrast in predictive success between the higher and lower end of the predictability scale, which disconfirms our second hypothesis.

However, before we discard our second hypothesis altogether, a number of issues need to be clarified. First, it may be the case that there are just too few instances to get reasonable predictions for gender and de/het. Second, monomorphemes are hardly an unbiased sample from the complete noun lexicon, and this may have its impact on classification accuracy as well. Third, given the apparent parallellism between low predictive accuracy (table 4) and low information value of attributes (table 2), the choice of attributes may be a crucial factor. We will take up these issues below.

4.3 Experiment 2

When we reformulated the lexical processes of diminutive formation, main stress assignment and gender assignment as classification tasks, two series of data sets were created, one comprising 5000 monomorphemic words, and one containing 15000 noun lemmas, either monomorpemic or complex. So far we have dealt only with the former. In the discussion of the results for monomorphemic words questions were raised concerning the bias implicit in restricting the noun lexicon to monomorphemic words, and keeping the sample size fixed. Both of these issues will be taken up in the next experiment.

The basic data set for the second experiment contains 15000 noun lemmas, which are either monomorphemic or complex. From this set, 15 samples were taken, varying in size from 1000 to 15000 items, with a step size of 1000. These data sets were encoded in the same manner as the monomorphemic nouns in experiment 1, i.e. 16 predictor attributes based on a syllabified phonological transcription, and class attributes depending on the application domain. For a full description of the encoding scheme, the reader is referred back to section 2.2. Classification experiments were run for all of these data sets using IB1-IG, using a ten-fold cross-validation methodology.

Before we discuss the results of the experiments, it is useful to look at a single sample in some detail, in order to compare its properties to those of the monomorphemic data sets. In table 5 figures for the entropy, average mutual information, equivalent number of attributes and noise-to-signal ratio are given for each of the four domains, using the same sample size (5000 items) as for the monomorphemes.

	$H(C)$	$\overline{M}(C,X)$	$EN.attr$	$NS.ratio$
diminutive	1.71	0.273	6.27	11.02
stress	2.04	0.343	5.93	8.57
gender	1.99	0.143	13.96	22
de/het	0.79	0.029	26.94	110.05

Table 5: Data analysis (5000 items)

When we compare this table to tables 1 and 2, it becomes apparent that lifting the restriction to monomorphemes affects the class distributions for all four data sets. In general, entropy is somewhat higher, with a 0.1 increase for de/het and a 0.2 increase for diminutives and gender. For stress, however, entropy is almost doubled, raising the effective number of classes from 2.13 to slightly less than 4. This underscores the distinction made in section 2.1 between the principles governing stress for monomorphemes and those for stress in complex words: for monomorphemes, the s-4 class had only marginal status, containing just a handful of exceptions to the three-syllable window generalization. For compound stress, main stress falls on the first member, which raises the number of representatives of class s-4 substantially.

The shift in category distributions has different implications for each of the

individual domains. For diminutives, the impact seems negligable, considering that the average mutual information of the attributes remains about the same. For stress, the mutual information value rises by 50%, reflecting the greater importance of syllables farther removed from the righthand word-edge. For both gender and de/het, however, there is marked increase in average mutual information, and a corresponding decrease in noise-to-signal ratio and equivalent number of attributes. For gender, the latter quantity now approaches the actual number of attributes supplied.

The impact of these changes on the classification results can be seen from table 6:

	accuracy	I_a	I_r
diminutive	99.20%	1.68	98%
stress	81.66%	1.46	71%
gender	82.80%	1.57	78%
de/het	92.30%	0.60	75%

Table 6: Results for Experiment 2 (5000 items)

The first thing to note is a huge increase in predictive performance for the 'arbitrary' categories. For the monomorphemes, accuracy on these tasks was low, and the relative information scores revealed that little (if anything) was learned about the domain. Compared to those results, accuracy rates have risen considerably (from 73.32% to 92.30%) for de/het, and are more than twice as high (82.80% vs. 40.72%) for gender. Even more importantly, the relative information scores now indicate that substantial headway has been made into solving these classification tasks, despite the fact that class entropy was higher for both tasks, and that no changes have been made to either the size of the data set or to the attributes used. For stress, the introduction of compounds and derivations into the data sets had the most profound effect: from a theoretical perspective, different principles are operative in stress assignment for monomorphemes, compounds and derivations. From a classification perspective, we saw that the class entropy was effectively doubled, and that major shifts occurred in the predictive value of individual attributes. This is reflected in the figures for accuracy and relative information score, both of which have dropped by about 10%. For diminutives, the situation is similar to the one observed for gender and de/het, albeit on a smaller scale. This problem was learned almost perfectly for the monomorphemes, leaving little room for improvement. Yet, despite the higher class entropy, there is a slight increase in both accuracy and relative information score.

When we compare the results for the different problem domains to each other, we see that performance for gender and de/het is now comparable to (or even better than) that for stress. There is however still a gap in performance between the predictions for diminutives and those for the other categories.

In the discussion of Experiment 1, we also brought up the issue of sample size.

To assess the impact of sample size on predictive accuracy, we set up Experiment 2 as a learning curve experiment. From the total data set containing 15000 items, progressively larger samples were taken, starting with 1000 items. In each subsequent sample, sample size was augmented by 1000. When we plot the accuracy obtained on each sample, we get the curve in figure 1:

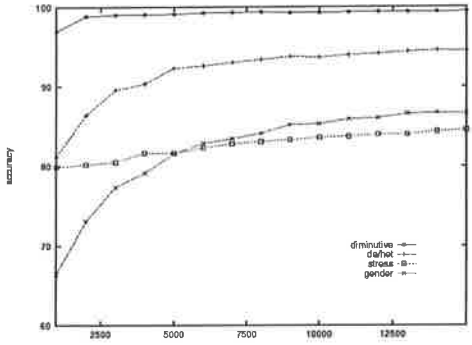

Figure 1: Accuracy from 1000 to 15000

From this figure, it can be seen that the sample we singled out above does not represent the highest accuracy achievable. For all problems, except for diminutives, there is still considerable improvement after this point; although the curves flatten out at about 8000 items, accuracy keeps increasing, even at 15000 items. At the end of the curve, the performance gap between diminutive and the other categories has narrowed, but it hasn't been closed. The question whether this can be achieved eventually, will remain unanswered for now.

A closer look at the results for stress, however, may provide some hints about future directions to be taken. As we have seen, extending the data sets with compounds and derivations, increased the complexity of the domain substantially. During the discussion, we have related this to the linguistic analysis, which posits different underlying principles for monomorphemes, compounds and derivations. For both compounds and derivations, morphological structure is assumed to play an important role, whereas for monomorphemes, the analysis can be stated entirely in phonological terms. For our experiments, we only relied on phonological features; the observed drop in accuracy for stress in Experiment 2 may well be due to the fact that crucial morphological features were missing from the encoding. For gender too, enriching the feature set with semantic and morphological information might improve predictive accuracy.

5 Conclusion

At the outset of our paper, we formulated two hypotheses concerning the predictability/arbitrariness of lexical categories. The first was that the degree to which

a lexical category is predictable can be shown quantitatively by standard data-analysis and machine learning techniques. This hypothesis was explored by examining four classification tasks, covering three different domains. The results of this analysis were found to be in close agreement with linguistic intuitions about the predictability of these domains. Our second hypothesis was that Memory-Based Learning succeeds in succesfully learning lexical categories, irrespective of their place on the predictable/arbitrary continuum. This hypothesis was tested by machine learning experiments with the IB1-IG classifier. We found clear indications for learnability in all domains, but varying degrees of success.

References

Aha, D. W.(1990), *A Study of Instance-based Algorithms for Supervised Learning Tasks*, PhD thesis, University of California, Irvine, CA.

Aha, D. W.(1992), Generalizing from case studies: a case study, *Proceedings of the Ninth International Conference on Machine Learning*, Morgan Kaufmann, San Mateo, CA, pp. 1–10.

Booij, G. E.(1995), *The Phonology of Dutch*, Oxford University Press, Oxford.

Daelemans, W. and van den Bosch, A.(1992), Generalization performance of back-propagation learning on a syllabification task, *in* M. Drossaers and A. Nijholt (eds), *Proceedings of TWLT3: Connectionism and Natural Language Processing*, Twente University, Enschede, pp. 27–37.

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

Daelemans, W., van den Bosch, A. and Zavrel, J.(1999), Forgetting Exceptions is Harmful in Language, to appear in Machine Learning.

Deutsch, W. and Wijnen, F.(1985), The article's noun and the noun's article: Explorations into the representation and access of linguistic gender in Dutch, *Linguistics* **23**(5), 793–810.

Gussenhoven, C. and Jacobs, H.(1998), *Understanding Phonology*, Arnold, London.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and van den Toorn, H. C. (eds)(1997), *Algemene Nederlandse Spraakkunst*, 2 edn, Wolters Plantyn, Deurne.

Henery, R. J.(1994), Methods for Comparison, *in* D. Michie, D. J. Spiegelhalter and C. C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, NY, chapter 7, pp. 107–124.

Kager, R. W. J.(1989), *A Metrical Theory of Stress and Destressing in English and Dutch*, PhD thesis, Rijksuniversiteit Utrecht, Utrecht.

Kononenko, I. and Bratko, I.(1991), Information-Based Evaluation Criterion for Classifier's Performance, *Machine Learning* **6**, 67–80.

Neijt, A. and Van Heuven, V.(1992), Rules and Exceptions in Dutch stress, *in* R. Bok-Bennema and R. Van Hout (eds), *Linguistics in the Netherlands*, Benjamins, Amsterdam, pp. 185–196.

Trommelen, M. T. G.(1983), *The Syllable in Dutch*, Foris, Dordrecht.
van Berkum, J. J. A.(1996), *The psycholinguistics of grammatical gender*, PhD thesis, Max Planck Instituut voor Psycholinguïstiek, Nijmegen.
van den Bosch, A.(1998), Careful Abstraction from Instance Families in Memory-Based Language Learning, to appear in JETAI.