Predicting Grammatical Classes from Phonological Cues: An Empirical Test [*]

Gert Durieux and Steven Gillis
Center for Dutch Language and Speech - CNTS
Department of Linguistics - GER
University of Antwerp - UIA

Abstract

This paper investigates to what extent the grammatical class(es) of a word can be predicted on the basis of phonological and prosodic information only. We report on several experiments with an artificial learning system which has to assign English wordforms to their appropriate grammatical class, using various types of phonological and prosodic information. First of all, we examine several phonological cues which were claimed by Kelly (1996) to be particularly good for distinguishing English nouns from verbs. Our results indicate that these cues are indeed partially predictive for the problem at hand and reveal that a combination of cues yields significantly better results than those obtained for each cue individually. We then show experimentally that 'raw' segmental information, augmented with word stress, allows the learning system to improve considerably upon those results . Secondly, we investigate several generalizations of the approach: basic segmental information also proves to be more predictive when the task

[*] Preparation of this paper was supported by a VNC project of FWO - NWO (contract number G.2201.96) and by a GOA grant (contract number 98/3). Thanks are due to Walter Daelemans, Frank Wijnen and the participants in the TROPICS conference for interesting discussions of phonological bootstrapping, and to Annick De Houwer for her critical reading of the manuscript.

is extended to encompass all open class words in English, and these findings can be replicated for a different (though related) language such as Dutch.

1. Introduction

How do children learn the (major) form classes of their language? How do they learn that "*table"* is a *noun*, "*nice"* an *adjective*, and "*kiss"* a *verb* as well as a *noun*? Form classes may be part of children's innate linguistic knowledge, which implies that a child "knows" that in the language (s)he is exposed to there are nouns and verbs, etc. However, just knowing that there are specific form classes is not sufficient, as this still leaves the problem of how the child determines which words in the speech stream belong to which class. One solution is to hypothesize that the child's knowledge about form classes includes procedures for discovering the (major) form classes in the language. If those procedures are part of the child's native endowment, they would have to consist of universal surface cues signaling grammatical category membership. At present it is unclear if such universally valid cues exist, and, if so, how they are to be characterized.

An alternative solution is that the child uses information that "correlates" with form class and finds a bootstrap into the system of formal categories. Several bootstrapping approaches have been proposed:

- *Semantic bootstrapping*: Under this approach, the meanings of words are used as a basis for inferring their form class (see e.g. Pinker 1987, Bates & MacWhinney 1989). In this line of thinking, Gentner (1982) noted that, as object reference terms, nouns have a particularly transparent semantic mapping to the perceptual/conceptual world, and children may use this mapping to delineate the category of "nouns".
- *Syntactic* (also *correlationa*l or *distributional*) *bootstrapping*: This approach holds that grammatical categories can be discovered on the basis of distributional evidence (see e.g. Maratsos & Chalkley 1980, Finch & Chater 1992, Mintz, Newport & Bever 1995). For instance, Mintz et al. (1995) show that by monitoring the immediate lexical contexts of words, the similarity of those contexts can be used to cluster lexical items and that the resulting clusters coincide with grammatical classes. More specifically, in

an analysis of the lexical co-occurrence patterns, Mintz et al. show that a window of one word to either side of the target word is sufficient to identify nouns and verbs.

- *Prosodic* and *phonological bootstrapping*: This approach holds that there are phonological and prosodic cues that may point the child to specific linguistic structures, e.g. clauses and phrases or specific classes of words, such as "open" vs. "closed" class words, "lexical" vs. "functional" items, or specific grammatical form classes (see e.g. Gleitman, Gleitman, Landau & Wanner 1988, Morgan, Shi & Allopena 1996).

All of these bootstrapping approaches typically emphasize the use of information from one domain, the "source" domain, to break into another domain, the "target" domain, and may thus be labeled "inter-domain" bootstrapping approaches, in contrast to the recently introduced notion of "autonomous" bootstrapping, which applies within a single domain (Cartwright & Brent 1996). In addition, it is typically argued that there is only a partial, i.e., a non-perfect correlation between the source and the target domains.

2. Phonological Cues to Syntactic Class

The usefulness of semantic and syntactic/distributional information is firmly established in the literature on grammatical category acquisition and assignment. The usefulness of phonological information is less straightforward. In a recent survey article, Kelly (1996) adduces several reasons why this may be the case: on the one hand, phonological cues are likely to be language-specific, and thus cannot be known in advance by the learner. By contrast, mappings between semantic and grammatical classes are assumed to be universal and may provide a useful bootstrap if the learner expects such mappings to occur. On the other hand, syntactic criteria remain the ultimate determinants of grammatical classes; any correlating phonological information can, at best, be supportive when both agree, or downright misleading when they do not. Hence, phonological cues are largely neglected as rare, unnecessary, unreliable and language-specific.

Still, in Kelly (1992, 1996) a large body of evidence is presented in support of the claims that phonological correlates to grammatical classes do exist for English, and

that people are sensitive to these correlates, even if they are only weakly diagnostic of grammatical classes. A fairly reliable correlate seems to be found in the stress pattern of disyllabic words: an examination of 3,000 disyllabic nouns and 1,000 disyllabic verbs, drawn from Francis & Kucera (1982), revealed that 94% of nouns have a trochaic (initial) stress pattern, whereas 69% of verbs display an iambic (final) stress pattern. More importantly, 85% of words with final stress are verbs and 90% of words with initial stress are nouns. Subsequent experiments in which subjects either had to construct sentences with a disyllabic word which could have either stress pattern, or read target sentences containing a disyllabic non-word in either nominal or verbal position, showed an outspoken preference for linking iambic words with the verb category and trochaic words with the noun category. These findings clearly indicate that a word's phonological form cues its grammatical class and that speakers appear to be sensitive to those links between form and class.

Another cue mentioned by Kelly (1996) is the number of syllables in a word: nouns tend to have more syllables than verbs, even when inflectional suffixes are added to the stem. In a corpus of English parental speech, the observed probability that a monosyllable is a noun was 38%. For disyllabic words this figure went up to 76% and for trisyllabic words to 94%. All words of four syllables were nouns. In a subsequent experiment, adults judged ambiguous (i.e., between noun and verb) monosyllables to be used more often as a verb and trisyllabic words to be used more often as a noun, when in fact both usages were equally likely.

Other, less reliable, phonological correlates of the grammatical classes noun and verb in English include (a) duration, (b) vowel quality, (c) consonant quality and (d) phoneme number (Kelly 1996: 252): (a) nouns are generally longer than verbs (controlling for syllable number, acoustic measurements show that the duration of nouns is generally longer than that of verbs), (b) nouns have more low vowels, (c) nouns are more likely to have nasal consonants, and (d) nouns contain more phonemes (again controlling for syllable number). Although these latter cues are deemed less reliable, Kelly (1992) does not preclude the possibility that, taken together, these individually weak cues may prove to be strong predictors of grammatical class.

This brings us to a delineation of the specific research questions addressed in this paper. Assuming the existence of phonological correlates to grammatical classes

and people's sensitivity to them, we want to explore their potential value as predictive cues for category assignment, given a lexicon of reasonable size. This exploration will proceed by running machine learning experiments which involve the relevant cues. The specific machine learning algorithm used is introduced in Section 3.

In Section 4 we test the predictive power of the phonological cues identified by Kelly (1996). Our first objective is to assess the predictive value of stress. In order to do so, we will report on a series of experiments covering increasingly larger segments of the English lexicon. A first test only includes disyllabic homographs with both a noun and a verb reading. A second test includes a larger number of disyllabic words, not all of which are necessarily ambiguous, and a third test includes a random sample of words containing one to four syllables. These experiments will allow us to assess whether the applicability of stress as a cue is restricted to the observed subregularity within the English lexicon, or whether it extends into broader regions of lexical space. A second objective is to assess the value of the "less reliable cues". To this end, we have constructed a test for each of the observed minor cues, using a setup similar to the final test for stress. A related focus of interest is the question whether these cues prove more predictive when used in combination.

A third objective is to determine whether the phonological makeup of words restricts grammatical class without *a priori* identification of the relevant cue(s). In all previous experiments, higher order phonological cues such as vowel height, nasals versus other consonants, etc., were coded into the learning material. Thus, it was taken for granted that the learner was able to extract these cues from the learning material or that the learner somehow had access to them. The final experiments reported in Section 4 drop this precondition from the experimental setup. More specifically, it is investigated to what extent the learner can detect the link between phonology and grammatical class when no a priori idenfication of the relevant phonological dimensions is performed and only syllabified strings of segments are supplied as learning material.

In Section 5 we will lift some of the restrictions observed in Section 4: the experiments reported on in this section will neither be restricted to English nor to the Noun/Verb opposition. In a first batch of experiments, we investigate to what extent the observed link between phonology and grammatical class carries over to another

language. Using Dutch as a testbed, we investigate whether Kelly's cues lead to reasonable results in predicting Dutch nouns and verbs, and whether the segmental material contains enough indications for distinguishing those categories in Dutch. In a second set of experiments, the task is broadened from predicting nouns and verbs in English and Dutch to predicting all open grammatical classes in those languages. Finally, we will deal briefly with issues of learnability and explore the relationship between the amount of training data the machine learner receives and its success in predicting grammatical classes. A related question concerns the impact of providing the learner with items of varying frequency: does the learner's performance on the task improve when high frequency items are provided? These latter issues form the bridge to the last section in which the crucial issue of the use of phonological bootstrapping as a language acquisition strategy will be considered in the light of the experimental results.

3. The Learning Algorithm

In this study we use a modified version of Instance-Based Learning (IBL, Aha, Kibler & Albert 1991). IBL is a 'lazy learner': no explicit abstractions such as rules are constructed on the basis of examples. This distinguishes 'lazy learning' from 'eager learning' approaches such as C4.5 (Quinlan 1993) or connectionist learning, in which abstract data structures are extracted from the input material (viz. decision trees in the case of C4.5, matrices of connection weights in the case of connectionist nets). IBL's learning consists of storing examples (or instances) in memory. New items are classified by examining the examples stored in memory and determining the most similar example(s) according to a similarity metric. The classification of that nearest neighbor (or those nearest neighbors) is taken as the classification of the new item. Thus IBL assumes that similar instances have similar classifications.

IBL falls within the class of supervised learning algorithms: the system is trained by presenting a number of input patterns (examples which coded as a vector of features) together with their correct classification. Testing the system consists in presenting previously unseen wordforms, suitably coded as feature vectors, and having the system predict their grammatical class. For the linguistic task in this study, viz.

grammatical class assignment, IBL's basic mode of learning is as follows: in the *training* or *learning* phase, precategorized items are presented to the system in an incremental way to the system. Thus, the system receives wordforms and their grammatical class. These examples (the *training items*) are stored in memory. In a *test* phase, the system carries out the required task. In this case, IBL has to predict the grammatical class of a novel wordform (a *test item*), i.e., a wordform not encountered during training. For this prediction the system relies on an explicit procedure for determining the similarity of the test item with the training items present in memory. IBL determines the most similar training item in its memory, and the grammatical category of that nearest neighbor is predicted to be the grammatical category of the test item.

The basic algorithm of IBL (Aha et al. 1990) determines similarity using a straightforward overlap metric for symbolic features: it calculates the overlap between a test item and each individual memory item on an equal/non-equal basis (see (1), where X and Y are examples or instances, and x_i and y_i are the values of the i-th attribute of X and Y) on an equal/non-equal basis (see equation 2):

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

$$(2) \qquad \delta(xi, yi) = 0 \ \ if \ \ x_i = y_i \ \ else \ \ 1$$

This similarity metric treats all attributes as equally important. Consequently, if there are irrelevant attributes, two similar instances may appear to be quite dissimilar because they have different 'unimportant' attributes. This is why we extended the basic algorithm with a technique for automatically determining the degree of relative importance of attributes. The basic idea is to modify the matching process of the test item with the memorized items in such a way that the importance of individual features is used in making the similarity judgment. In other words, features that are important for the prediction should be made to bear more heavily on the similarity judgment. A weighting function ($G(a_i)$) was introduced in equation (1), yielding equation (3).

$$(3) \qquad \Delta(X,Y) = \sum_{i=1}^{n} G(a_i)\, \delta(x_i, y_i)$$

The function computes for each attribute its *information gain* over the entire set of training items or memorized instances. This information theoretic measure is perhaps best known from Quinlan's work on the induction of decision trees (Quinlan, 1986, 1993).

The information gain of a particular attribute *a,* or in other words, the information gained by knowing the value of attribute *a*, is obtained by comparing the information entropy of the entire training set (H(T)) with that of the training set restricted to a known attribute *a* ($H_a(T)$). The gain of information is the difference between these measures as indicated in equation (4):

$$(4) \qquad G(a) = H(T) - H_a(T)$$

The entropy of the training set is computed using equation (5): the entropy of the training set equals the average amount of information needed to identify the class of a single instance and is computed as the sum of the entropy of each class in proportion to its frequency in the training set.

$$(5) \qquad H(T) = -\sum_{i=1}^{j} \frac{f(C_i)}{|T|} \log_2 \frac{f(C_i)}{|T|}$$

(where $f(C_i)$ is the frequency of class C_i in the training set and |T| is the total number of cases in the training set)

The entropy of the training set restricted to each value of a particular attribute is computed in a similar way, i.e., the average information entropy of the training set restricted to each possible value of the attribute is calculated using (5). As expressed in equation (6), the weighted sum of these entropy measures yields the expected information requirement.

$$(6) \qquad H_a(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

The information gain of an attribute (see (4)) expresses its relative importance for the required task. Used as a weighting function (as expressed in (3)) in determining similarity, attributes will not have an equal impact on determining the nearest neighbor of a test item: instances that match on important attributes (attributes with a high information gain value) will eventually turn out to be nearer neighbors than instances that only match on unimportant attributes (attributes with a low information gain value). (For a more extensive discussion of our implementation of IBL, we refer to Daelemans & van den Bosch 1992, Daelemans et al. 1994, Gillis et al. In press).

4. Phonological cues for English Nouns and Verbs

The experiments reported in this section are meant to investigate how accurately English wordforms can be assigned to the classes Noun and Verb. First, the phonological cues identified by Kelly (1996) will be used in machine learning experiments in order to assess their predictive power. Next, alternative phonological cues - as represented in the training material provided to the algorithm -will be explored and their strength will be compared to that of Kelly's phonological cues.

4.1. Data and Method

All data for the experiments were taken from the CELEXv2 lexical database (Baayen, Piepenbrock & van Rijn 1991). This database was constructed on the basis of the Collins/Cobuild corpus (17,979,343 words), which was compiled at the University of Birmingham and augmented with material taken from both the Longman Dictionary of Contemporary English and the Oxford Advanced Learner's dictionary.

The whole lexical database comprises 160,595 wordforms, belonging to 52,447 lemmas. For the experiments, we restrict the database to nouns and verbs encountered

at least once in the Collins/Cobuild corpus. We shall refer to this restricted database of nouns and verbs simply as 'the database'.

All experiments were run using the 'leaving-one-out' method (Weiss & Kullikowski 1991) to get the best estimate of the true error rate of the system. In this setup, each item in the dataset is in turn selected as the test item, while the remainder of the dataset serves as training set. This leads to as many simulations as there are items in the dataset: in each simulation the entire dataset is used for training, except for one item which is used for testing. The success rate of the algorithm is obtained by simply calculating the number of correct predictions for all words in the test set.

4.2. Experiment 1: Stress

In a first experiment, we investigate IBL's ability to predict grammatical class using the stress pattern of wordforms. Kelly (1996) claims that the large majority of disyllabic nouns are trochees while a majority of disyllabic verbs are iambs. For wordforms such as *"abstract"*, which are orthographically ambiguous between a noun and a verb reading, not a single pair exists where the noun has iambic stress while the verb has trochaic stress. The experiment was set up to test Kelly's claim that stress is a good predictor of grammatical class and to test the generality of that claim. For this purpose three datasets were constructed. The first dataset was restricted to orthographically ambiguous disyllabic words of the type *"abstract"*. The second dataset was compiled from all disyllabic wordforms in the database, lifting the restriction that the noun and the verb should be orthographically identical. The third dataset was a selection from all noun and verb wordforms in the database. Enlarging the dataset in this way will allow us to assess the predictive value of stress and to assess the generality of its predictive power.

The first dataset consists of all disyllabic orthographical doublets found in the database (henceforth: "Disyllabic Homographs"). This dataset contains 212 nouns, 215 verbs and 16 ambiguous wordforms. Each of these wordforms is coded using two features, corresponding to the stress level of its syllables. In the encoding we use "2" to denote primary stress, "1" to denote secondary stress and "0" to indicate that the

syllable bears no stress. The target categories, i.e., the grammatical classes, are coded as "N" for *noun*, "V" for *verb*, and "NV" for ambiguous wordforms. For instance, in the training set, the word "*abstract*" is represented as the triple <2,0,N> for the noun reading and <0,2,V> for the verb reading. The first element in the triple denotes the stress pattern of the first syllable, the second element the stress pattern of the second syllable, and the third element denotes the target category, viz. the grammatical class of the word. This means that in the learning phase the algorithm encounters the wordform "*abstract*" twice, once as the pattern <2 0> with its target category <N>, and once as the pattern <0 2> with its associated target category <V>. The wordform "*uses*" is presented only once, viz. as the pattern <2 0> and its associated target category <NV>.

The results in Table 1 indicate that the stress pattern strongly constrains the possible grammatical categories: solely on the basis of the stress pattern, the grammatical category can be accurately predicted in 82.6% of the cases. The number of correctly predicted nouns and verbs is almost identical: in both cases the success rate exceeds 84%. Not surprisingly, wordforms which are phonologically indistinguishable, such as "*being*", are very poorly predicted (NV in Table 1). Although Kelly's observation that not a single homograph exists where the noun has iambic stress and the verb trochaic stress applies to this dataset as well, this does not imply that stress makes perfect predictions. First of all, not all iambic wordforms are verbs, and not all trochaic wordforms are nouns: in wordforms such as "*uses*", the difference between the noun and verb reading lies in the (lack of) voicing of the first "s", not in the stress pattern. In words such as "*cashiers*" the difference lies in the first vowel, which is reduced to schwa under the verb reading. Second, the presence of phonologically indistinguishable wordforms in the dataset considerably complicates the prediction task.

INSERT TABLE 1 ABOUT HERE

These results indicate that word stress is a good predictor of a word's grammatical class, provided that we restrict the dataset to disyllabic nouns and verbs which are orthographically ambiguous homographs (such as "*abstract*"). How general is this finding, or in other words, how robust is stress as a cue for predicting that a given wordform is a noun or a verb? For this purpose we expanded the dataset to (a) a random

selection of all disyllabic words, and (b) a random selection of all wordforms of the database.

(a) The dataset was expanded to include other disyllabic words than homographs (henceforth: "Disyllabic Wordforms"). Since these are far more numerous, a random stratified selection of 5,000 items was made, consisting of 3,142 nouns (62.84%), 1,465 verbs (29.3%) and 393 ambiguous wordforms (7.86%).

(b) The dataset was expanded to include wordforms of up to four syllables (henceforth: "All Wordforms"): we selected a random stratified sample of 5,000 items from the database containing 3,136 nouns (62.72%), 1,457 verbs (29.14%) and 407 ambiguous wordforms (8.14%). Since wordforms were no longer of equal length, the coding scheme had to be adapted slightly: as in the previous experiments, we used one feature per syllable, indicating the syllable's stress level (i.e., primary, secondary or no stress). Words containing fewer than four syllables were padded to the left with null features ("-"). This implies that wordforms are aligned to the right, which is consistent with current analyses in metrical phonology where stress in English is assigned from right to left.

Thus, in the Disyllabic Homographs data "*abstract*" is represented as the triples <2,0,N> and <0,2,V>. In the Disyllabic Wordforms data "*wordform*" is represented as <2,0,N>. In the All Wordforms data "*phonology*" is represented as quintuple <0,2,0,1, N>, in which the first four values represent the stress level of the first through the fourth syllable and the fifth element represents the target category of the word. A wordform with fewer than four syllables such as "*wordform*" is represented as the quintuple <-,-,2,0,N> in which the first two values represent empty slots, the third value represents the stress level of the prefinal syllable, the fourth value the stress level of the final syllable and the fifth value the target category of the word.

Table 2 displays the results of the learning experiment with these two new datasets. In comparison with the Disyllabic Homographs (overall success score: 82.6%), the overall success scores for the Disyllabic Wordforms and All Wordforms are far inferior: 69.74% and 66.18% respectively. This drop in accuracy is most spectacular for verbs: whereas in the Disyllabic Homographs dataset verbs were correctly classified in almost 85% of the cases, this level of accuracy drops to 38.9% (Disyllabic Wordforms) or less (All Wordforms). In both cases, verbs were erroneously classified

as nouns. The ambiguous NV category is never predicted correctly: apparently, stress is not an accurate diagnostic for this category.

INSERT TABLE 2 ABOUT HERE

Taken together, these results show that stress is a good predictor in the case of disyllabic homographs, but already far less reliable a predictor when all disyllabic wordforms are taken into account. When a still larger fragment of the lexicon is considered, the predictive value of stress further diminishes. It seems then that Kelly's characterization of stress as a reliable cue needs serious qualification: only in the case of disyllabic homographs can stress be labeled "reliable" as a predictor of grammatical class. For larger portions of the lexicon, the value of stress seems rather dubious. This conclusion is further strengthened when we study the Information Gain (see Section 3 for a formal definition) of the feature Stress in all three datasets. The Information Gain of stress is plotted in Figure 1 for the Disyllabic Homographs, Disyllabic Wordforms and All Wordforms datasets (restricted to the values for the final and the prefinal syllable).

INSERT FIGURE 1 ABOUT HERE

The Information Gain values for stress show a very clear picture: the value for Disyllabic Homographs is very high in comparison with the values for Disyllabic Wordforms and All Wordforms. This means that there is a high gain of information when the stress pattern of the word (at least the stress level of the two last syllables) is known in the case of the Disyllabic Homographs, while for the two other datasets the gain of information is far less. This difference in Information Gain value explains why the prediction of grammatical classes for Disyllabic Homographs is significantly better than the prediction for Disyllabic Wordforms and All Wordforms. Stress is simply a far worse predictor in the latter two conditions than in the former condition.

4.3. Experiment 2: Less reliable cues

In addition to stress, Kelly (1996) identifies a number of “less reliable cues”: nouns (a) are generally longer than verbs (controlling for syllable number), (b) have more low vowels, (c) are more likely to have nasal consonants, and (d) contain more phonemes (again controlling for syllable number). In order to test the predictive value of these cues, we set up machine learning experiments similar to the ones in the previous section. Each of the cues was tested separately and an experiment with a combination of the cues was run. More specifically, the experiments cover vowel height (b), consonant quality (c) and number of phonemes (d). The first cue, duration (a), was not covered in our experiments, since the CELEX lexical database does not contain acoustic measurements that would allow a suitable encoding.

For the sake of comparison, we use the same random stratified sample of 5,000 words in all the experiments we report on in this section, which is identical to the one the All Wordforms dataset from the previous section was derived from. The sample contains 3,136 nouns, 1,457 verbs and 407 ambiguous wordforms. Word length varies from one to four syllables. We first describe the actual encoding of the various “less reliable cues” and then present a global overview of the results and a detailed comparison of the success scores.

(a) Vowel Height: Each syllable of a wordform is coded for the vowel height of the nucleus. One feature per syllable is used, indicating vowel height of the syllable nucleus. Values for this feature are "high" (for the vowels /ˆ, i:, ¨, u:/), "mid" (for the vowels /´, 3:/) and "low" (for the vowels /æ, ø:, Å, a, √/). For diphthongs, a difference is made between “closing” diphthongs, which involve tongue movement from mid or low to high and “centering” diphthongs, where movement occurs from a peripheral to a central position. “Closing” diphthongs are /eˆ, aˆ, øˆ, \¨, a¨/, and “centering” diphthongs /ˆ\, ´\, ¨\/. For schwa and syllabic consonants the (dummy) feature value “neutral” was used. Words containing fewer than four syllables were padded to the left with null features.

(b) Consonant Quality: For the second experiment wordforms are coded for consonant quality. Here, two features per syllable are used, indicating the presence ("true") or absence ("false") of nasals in the onset and the coda of the syllable. As in the

previous experiment, shorter wordforms are left-padded with null features, yielding a total of eight features.

(c) Number of Segments: For the third experiment, four features are used, indicating the number of segments per syllable.

(d) Combined Cues: For the fourth experiment all these cues were combined: wordforms were coded for vowel height, consonant quality, number of segments and stress (the latter as explained in previous section).

Table 3 shows the results of these experiments: in the second column the success rates for the Stress encoding are displayed (see Table 2 in previous section), followed by those for Vowel Height, Consonant Quality, and Number of Segments. The last column contains the success scores for the combination of these cues (Combined Cues).

INSERT TABLE 3 ABOUT HERE

The success scores in Table 3 show that taken individually, Stress, Vowel Height, Consonant Quality and Number of Segments are good predictors for the category of nouns. They are poor cues for verbs and completely non-diagnostic for the ambiguous noun/verb category. An analysis of the global success scores ('Total' in Table 3) reveals that Stress is not a significantly better cue than Vowel Height (χ^2 = 1.623, $p < 0.2027$) or Consonant Quality (χ^2 = 1.928, $p < 0.1649$). The result for the Number of Segments is significantly different from that for Stress (χ^2 = 15.397, $p < 0.0001$). Consequently, Kelly's (1996) characterization of Stress as a robust cue and the three other cues as fairly weak ones is clearly contradicted by the outcome of these experiments.

A second finding established by the total success scores in Table 3 is that, as hypothesized by Kelly (1996), a combination of the cues predicts the grammatical class of a word significantly better than any single cue in isolation (see the column Combined Cues in Table 3). On the basis of a combination of the four cues used in the experiment, the grammatical class of a word can be predicted with an accuracy of more than 67%. This score is significantly better than the score for Stress (χ^2 = 6.552, $p < 0.0105$), Consonant Quality (χ^2 = 15.587, $p < 0.0001$) and Number of Segments (χ^2 = 42.022, p

< 0.0001). However, there is no significant difference between the predictive power of Vowel Height and that of the Combined Cues ($\chi^2 = 1.654$, $p < 0.1984$). We will come back to this result later.

A third outcome is that no single cue is especially powerful in predicting a particular grammatical class. When we compare the accuracy of the predictions for the individual grammatical classes, all cues are far better in predicting nouns than in predicting the other two categories. It is not the case that a particular cue is especially reliable in predicting one category and another in predicting another category. Irrespective of what cue is used, the success score for nouns is higher than that for the other categories. Nouns can be most accurately identified: up to 94% of the nouns were correctly classified. Verbs gain most from a combination of the individual cues: the best score for the individual cues is 34% for Vowel Height, which is significantly less than the 61.37% for the Combined Cues. This means that a combination of the cues appears to be necessary for distinguishing nouns and verbs: when the cues are combined the success score for verbs attains a level approaching that for nouns. And herein lies the difference between the cue Vowel Height and the Combined Cues: their global success scores do not differ significantly, but verbs cannot be identified with any reasonable accuracy on the basis of Vowel Height alone, while this does appear to be the case when the cues are combined. The high success score for nouns (around 90%) in the single cue conditions and the decrease of the success score for nouns in the Combined Cues condition, taken together with the increase of the success score for verbs in that condition, suggests that individual cues are insufficient for distinguishing the categories of nouns and verbs, while differentiation takes place when the cues are combined. In the Combined Cues condition, the ambiguous noun/verb category cannot be distinguished from the other categories on the basis of the cues selected.

In conclusion, the results reported in this section indicate that there is more than an arbitrary relationship between English nouns and verbs and their phonological form. If our artificial learner had only made an 'educated guess', such as always predicting the most frequent category, a success rate of 62.72% (the percentage of nouns in the dataset) was to be expected. The mere fact that IBL reaches a score of more than 67% is suggestive of a closer link between grammatical classes and their phonological form. However, it may be argued that this conclusion is heavily biased because Kelly's *a*

priori analysis informed the coding of the learning material. In other words, IBL's task may have been simplified because (the) relevant phonological cues were 'precoded' in the learning material so that the learner 'knew' in advance what information was relevant for the task at hand and this knowledge may have guided the discovery of the relevant cuts between grammatical classes. This issue will be taken up in the next experiment: if we do not provide the relevant phonological cues and give the learner only a string of phonemes as input, can the learner discover the relevant phonological cues for assigning words to their grammatical class?

4.4. Experiment 3: Phonological encoding

In Experiment 2 all encodings were obtained by extracting specific features from the syllabified phonological string: in the Vowel Height encoding, syllabic nuclei were examined for one particular dimension, viz. vowel height. In the Consonant Quality encoding, the same was done with onsets and codas for the dimension nasality. The results for the Combined Cues indicate that considering more than one dimension at the same time gives rise to more accurate predictions. To investigate the extent to which relevant dimensions are picked up by the learner without their *a priori* identification, we set up an experiment in which the 'raw' phonological material is used.

In order to asses the importance of the cues identified by Kelly (1996) these cues are now contrasted to a plain segmental encoding. If the *a priori* cues define the only relevant dimensions, we expect equal or poorer performance from the segmental encoding. Equal performance would indicate that the phonological *a priori* cues provide all the information necessary for category assignment, and that substitution of these cues by the segmental material from which they were derived, does not add any relevant information. Poorer performance may occur when the relevant oppositions, as defined by the cues, are obscured by the introduction of other, less relevant aspects of the segmental material. In this case, the learning algorithm would simply be unable to uncover the important dimensions. If, on the other hand, the *a priori* cues do not exhaust all relevant oppositions, the impact on the results may go in either direction. In the worst case, a scenario similar to the one described above may occur: although all

potentially useful information is somehow present in the encoding, the algorithm is unable to single out the relevant dimensions, and performs less well than it should. If, however, the algorithm is capable of capitalizing on the extra information supplied in the segmental encoding, equal or better performance is expected. Equal performance would then indicate that other oppositions, although relevant, do not enhance overall prediction accuracy. Better performance would indicate that the cues as identified by Kelly (1996) leave out some important dimensions, which adversely effects the algorithms success rate.

In Experiment 3 the same random stratified data set of 5,000 items as in Experiment 2 was used to facilitate comparison of results. The dataset was coded as follows:

(a) For the first encoding, three features per syllable were used, corresponding to the onset, the nucleus and the coda (henceforth: ONC), yielding twelve features per word. As in the previous experiments, words containing less than four syllables were left-padded with null features.

(b) For the second encoding, stress was added to the ONC encoding (henceforth: ONC + Stress), to allow comparison with Experiment 1.

(c) For the third encoding, the Combined Cues of Experiment 2 were added to the ONC encoding (henceforth: ONC + Combined Cues).

The results of the experiments are displayed in Table 4. The total success scores range from 73.86% for ONC, 78% for ONC + Stress to 78.16% for ONC + Combined Cues. For the sake of comparison, Table 4 also includes the success scores of the Combined Cues (Experiment 2, see Table 3). A comparison of the ONC encoding with the Combined Cues encoding reveals that a plain phonemic representation of the wordforms results in substantially superior predictions. The global success score of the "ONC" encoding (73.86%) is significantly better than the one for the "Combined Cues" (67.68%, $\chi^2 = 34.039$, $p < 0.0001$). This also holds for the success scores for the individual categories (N: 79.24 % vs. 77.59%; V: 75.7% vs. 61.37%; NV: 25.8% vs. 11.96%). Thus, significantly better results are obtained when the learning material is presented as a syllabified string of segments, which implies that the cues identified by Kelly do not provide all the relevant information, since in that case the encoding of the wordforms as strings of segments would not have resulted in a significant increase of

the success scores. Further qualitative analyses will have to reveal if indeed IBL uses cues similar to those established by Kelly and/or if the algorithm bases its predictions on (entirely) different information.

INSERT TABLE 4 ABOUT HERE

A second finding is that stress is indeed a relevant factor for predicting nouns and verbs: the success score for ONC, viz. 73.86%, increases to 78.04% ($\chi^2 = 23.914$, $p < 0.0001$) when in addition to the segmental material suprasegmental information (viz. word stress) is added to the encoding of the training material. On the other hand, adding in the other cues (viz. "vowel height", "consonant quality" and "number of segments") does not bring about a significant increase in accuracy (ONC + Stress vs. ONC + Combined Cues: 78.04 vs. 78.16, $\chi^2 = 0.021$, $p < 0.8846$). This observation indicates that those cues do not add any relevant information beyond that which IBL already uses in the ONC + Stress encoding. In other words, the higher level phonological information that these cues bring to the task of category assignment does not significantly affect performance, which implies that this information is, in fact, redundant with respect to the segmental encoding.

In conclusion, the experiments reported on in this section show that the best performance in grammatical class prediction is attained by presenting the 'raw' phonological facts to the learning algorithm, i.e., syllabified strings of segments and the stress pattern of the wordform. Adding higher level phonological information does not lead to significantly better predictions. The fact that using the *a priori* cues results in inferior performance indicates that in abstracting away from the actual phonological facts, important information for solving the task is lost.

Now that we have shown that there is a close link between the segments and the stress pattern of English nouns and verbs, two questions about the generalizability of this finding come to mind. First of all, can this link also be shown to exist for other languages and, if so, how tight is this link? Secondly, can the link be shown to exist for all open class words, i.e. also for adjectives and adverbs in addition to nouns and verbs? The following sections address these questions. After extending the approach to another

language, viz. Dutch, we proceed to extend the grammatical classes to be predicted to all open classes.

5. Generalization

In the experiments reported on in the previous section, the predictability of English nouns and verbs (and the ambiguous Noun/Verb category) was investigated. It was shown empirically that the 'raw' phonemic encoding supplemented with prosodic information (stress pattern) yields the highest success score and that adding higher level phonological information does not improve the success score significantly. In this section, we will investigate to what extent these findings can be generalized. First, it will be examined if similar results can be obtained for another language. Mainly due to the immediate availability of appropriate data in the CELEX database, Dutch was chosen for this purpose. Secondly, it will be examined if the predictability of grammatical classes can be extended to all open classes for English as well as Dutch.

5.1. How general are Kelly's phonological cues?

In the previous section, the use of phonological cues for grammatical class assignment was investigated. More specifically, we investigated Kelly's (1996) claim that there are cues with varying predictive power for the categories Noun and Verb in English. Kelly (1996) assumes that the phonological cues are likely to be language-specific. Intuitively plausible as this assumption may be, it seems to be contradicted by an investigation of Morgan, Shi and Allopenna (1996) and Shi, Morgan and Allopenna (1998). They investigated if various "presyntactic cues" (such as number of syllables, presence of complex syllable nucleus, presence of syllable coda, and syllable duration, to name only a few phonologically relevant cues used by Shi et al. (1998: 174)) are sufficient to guide assignment of words to rudimentary grammatical categories. Their investigation of English (Morgan et al. 1996), Mandarin Chinese and Turkish (Shi et al. 1998) shows that "sets of distributional, phonological, and acoustic cues distinguishing lexical and

functional items are available in infant-directed speech across such typologically distinct languages as Mandarin and Turkish." (Shi et al. 1998: 199). Thus it may well be the case that the cues identified by Kelly are crosslinguistically valid - be it to a different extent for each language. This finding would be in line with the findings of Shi et al. (1998: 169): "Despite differences in mean values between categories, distributions of values typically displayed substantial overlap."

In order to explore this issue, we conducted experiments similar to those reported in the previous section, but using Dutch wordforms instead of English ones. In Experiment 4, the cues identified by Kelly (1996) are used in a grammatical class assignment task involving Dutch nouns and verbs. In Experiment 5, the conclusion from Experiment 3 that predictions based on "raw" phonemic material yield better results than predictions based on higher level phonological cues will be tested on the Dutch material.

5.1.1. Experiment 4: Predicting grammatical classes in Dutch using phonological cues

The aim of this experiment is the same as that of the second experiment: we investigate IBL's ability to predict grammatical class using the phonological cues identified by Kelly (1996), viz. the stress pattern of a wordform, the quality of the vowels, the quality of the consonants and the number of phonemes (controlling for syllable number). These cues are represented in the training material in machine learning experiments. Cues are represented individually as well as in combination, as was the case for the experiments with English wordforms.

In Experiment 4, Dutch wordforms are used. All the Dutch data for the experiments were taken from the CELEX lexical database. This database was constructed on the basis of the INL corpus (42,380,000 word tokens) compiled by the Institute for Dutch Lexicology in Leiden. The whole database contains 124,136 Dutch lemmas and 381,292 wordforms.

A random stratified set of 5,000 wordforms was selected from the CELEX lexical database for the sake of this experiment. The sample contains 3,214 nouns (64.28%), 1,658 verbs (33.16%) and 128 (2.56%) of ambiguous wordforms. Word length varies from one to four syllables.

The encoding schemes for the training data mirror those used in Experiments 1 and 2 closely. For the cue "Stress" each wordform was encoded using four features, the value of which indicated the stress level of the syllable (primary or no stress). Since CELEX, unfortunately, does not code Dutch wordforms for secondary stress, this feature value was lacking from our encoding as well. For the cue "Vowel Height", each syllable was coded for a single feature corresponding to the syllable nucleus. Values for this feature were based on Kager's (1989) description of the Dutch vowel system: "high" (/i:, i::, u:, y:, y::, ˆ, ∏/), "mid" (/´:, œ:, Å:, e:, ø:, o:, ´, œ, ø/), "low" (/a:, å/), "diph" (/´ˆ, åu, œy/), "neutral" (/\/). Standard Dutch does not have syllabic consonants. For the cue "Consonant Quality" two features per syllable were used, one for the presence or absence of nasals in the onset and in the coda. For the cue "Number of Segments", the number of segments in each syllable was coded. Moreover, as was the case in the experiment with English wordforms, wordforms with fewer than four syllables were left-padded with null features so as to satisfy fixed length input required by the implementation of the algorithm.

Table 5 shows the results of these experiments. The first column shows the target categories. In the second column the success scores for the cue "Stress" are mentioned, followed by "Vowel Height", "Consonant Quality" and "Number of Segments". The last column contains the success scores for the combination of the cues.

INSERT TABLE 5 ABOUT HERE

The algorithm reaches a total success score that ranges from 58% for "Stress" to 75% for "Combined Cues". If we take as a base line the success score of the algorithm when it would always predict the most frequent category, i.e., the category Noun, the base line would be 64%. "Stress" and "Consonant Quality" score significantly below this base line ("Stress": $\chi^2 = 42.2832$, $p < 0.0001$; "Consonant Quality": $\chi^2 = 28.8183$, $p < 0.001$), whereas "Vowel Height" and "Number of Segments" score above the base line ("Vowel Height": $\chi^2 = 5.2483$, $p < 0.0221$; "Number of Segments": $\chi^2 = 7.0294$, $p < 0.0080$). The "Combined Cues" show a considerable increase of the success score, i.e., the success score for the combined cues is significantly higher than that for the individual cues (Stress - Combined Cues: X2 = 335.488, p <0.0001; Vowel Height -

Combined Cues: $\chi^2 = 295.920$, $p < 0.0001$; Consonant Quality - Combined Cues: $\chi^2 = 295.920$, $p < 0.0001$; Number of Segments - Combined Cues: $\chi^2 = 86.099$, $p < 0.0001$).

A comparison of the results for Dutch with those for English (see section 4.3) reveals some striking similarities. First of all, in both languages nouns appear to be much easier to predict than verbs on the basis of phonological cues. The ambiguous noun/verb category appears to be impossible to delineate, and hence, to predict. Secondly, a combination of the partially predictive cues yields a significantly better success score than all individual cues taken in isolation. This means that even the cues that do not seem to be very informative in isolation bring valuable information to solving the task when that information is combined with other information.

A third and very remarkable finding is that even though the cues were initially designed for *English* nouns and verbs, they yield a fairly good result in predicting *Dutch* nouns and verbs. The success score for Dutch is even better than that for English: 68% for English and 75% for Dutch. This suggests at least two interpretations. The cues described by Kelly (1996) may reveal more than mere idiosyncrasies of the English language: even though Dutch is typologically closely related to English, the cues identified for English also hold for Dutch. Further investigation of other languages could shed light on the question whether in the phonological system of languages of the world there are particular dimensions that correlate with particular grammatical classes. Alternatively, it may well be the case that the relationship between Dutch phonology and syntax (grammatical classes) is so transparent that even very weak cues allow for fairly good predictions.

The first possibility, viz. that there are phonological cues to grammatical class that transcend language idiosyncrasies, requires a comparison of languages that goes far beyond the scope of this paper. Nevertheless, the results for Dutch are very striking in the sense that, in the literature reviewed in the introductory section, there appeared to be a consensus that phonological cues do not qualify for anything more than language idiosyncratic tendencies. The experiments reported on in this section show that, even with cues defined for English, Dutch nouns and verbs can be predicted correctly in three out of four cases, a success score that is higher, incidentally, than the one obtained for English. This seems to suggest a more than language specific link between phonological and syntactic structure. At this point we do not want to suggest that

Kelly's cues qualify for universal validity. On the contrary, the only suggestion is - in line with Shi et al. (1998) - that it may be fruitful to start the quest for cues that link phonology and syntax in a more than idiosyncratic way. That Kelly's cues may be improved upon during such an undertaking, can be exemplified by a simple additional experiment we performed: if we reformulate the cue "Consonant Quality" (the presence or absence of nasals in the onset and coda of syllables) in terms of the cluster types occur in those positions, the success rates for both English and Dutch improve significantly: from 64.86% to 68.70% for English and from 59.06% to 72.20% for Dutch. In other words, simply taking into account what qualifies as a legal syllable onset or syllable coda permits the algorithm to predict grammatical class membership of 74.8% of the Dutch nouns and 71.2% of the Dutch verbs and 75.67% of the English nouns and 66.3% of the English verbs.

The second possibility hinted at above was that the link between phonology and parts of speech is simply more transparent in Dutch than in English. This brings us to the question how well nouns and verbs can be predicted in Dutch. In the experiments on grammatical class assignment in English, it was shown that using "raw" segmental material instead of the cues abstracted from the segmental material yielded significantly better results in terms of success scores. In the following experiment this segmental encoding of the training material was applied to Dutch wordforms.

5.1.2. Experiment 5: Phonological encoding of Dutch wordforms

Experiment 5 was designed to test if like the outcome of Experiment 3 involving English wordforms a phonological encoding of Dutch wordforms yields a better success score than the other encodings in which abstract features such as vowel height were used to code the learning material. The same random stratified data set of 5,000 items as in Experiment 4 was used. The data were coded in a similar way as indicated for the experiment with English wordforms (see section 4.4), viz. (a) an ONC encoding in which the segments in the onset, the nucleus and the coda of each syllable are taken as the values of the features Onset, Nucleus and Coda; (b) an ONC + Stress encoding in which in addition to the ONC encoding also the stress level (primary or no stress) of

each syllable is coded; and (c) an ONC + Combined Cues encoding in which in addition to the ONC encoding also the cues identified by Kelly were used.

The results of the experiment are displayed in Table 6. It appears that the three encodings yield a very similar total success score: 82% for the ONC encoding and 83% for the ONC + Stress and the ONC + Combined Cues encodings. The main finding of the experiment is that in more than 80% of the cases segmental material suffices to classify nouns and verbs appropriately. Only the difference between the ONC encoding and the ONC + Combined Cues encoding is statistically significant ($\chi^2 = 4.459$, $p < 0.0347$). This means that adding the suprasegmental feature stress to the ONC encoding does not lead to a significantly better prediction of the grammatical classes noun and verb in Dutch, in contrast to the results obtained for English. This may be due to the fact that the stress encoding for Dutch, which only indicates primary or no stress, is less informative than the one for English, which also indicated secondary stress.

The individual categories are identified quite accurately as well on the basis of the segmental material: nouns (N) and verbs (V) are classified correctly in approximately 85% of the cases. However, the ambiguous NV category hardly reaches 30%. In other words, IBL is able to detect the relevant cues for assigning nouns and verbs to their appropriate class in more than eight out of ten cases, which is well above chance level. The ambiguous words are hard to classify: there does not appear to be robust information in the segmental material for distinguishing nouns and verbs from the ambiguous noun/verb wordforms.

INSERT TABLE 6 ABOUT HERE

A second important finding is that a segmental encoding yields significantly better results than an encoding in terms of more abstract phonological features. The column Combined Cues in Table 6 is taken from Experiment 4 (see Table 5) in which the data were encoded with Kelly's cues. The success score of the ONC encoding, 82%, is significantly higher ($\chi^2 = 76.79$, $p < 0.0001$) than the success score for the Combined Cues encoding (68%). This finding replicates the one found for English: in both languages the segmental encoding yields significantly better results than the encoding in terms of more abstract phonological features.

A comparison of the results for English and Dutch reveals that the relationship between the phonological form of a word and its grammatical class is more transparent in Dutch than in English. The ONC encoding yields a success score of 74% for English and 82% for Dutch, and adding stress to the encoding increases the success score to 78% for English and 84% for Dutch.

5.2. Experiment 6: Predicting Open Classes in English and Dutch

In the experiments presented in the previous sections, only nouns and verbs were considered. We investigated to what extent segmental and suprasegmental information was sufficient to predict whether a particular word is a noun, a verb, or belongs to the ambiguous noun/verb category. In this section, we consider all open grammatical classes, viz. verbs, nouns, adjectives, and adverbs and all ambiguous categories (such as noun/verb, verb/adjective, noun/verb/adjective, etc.). In this sense, the experiments reported in this section are complementary to the ones reported by Morgan et al. (1996) and Shi et al. (1998) who show that there are phonological (as well as other) cues that make open class and closed class items (lexical and functional items) detectable in principle. In our experiments, we investigate if on the basis of phonological information a further differentiation of the open class items is possible in principle.

The data for these experiments were extracted from the CELEX lexical database. Again a random stratified sample of 5,000 items was selected for each language. Table 7 gives an overview of the frequency distribution of the different word classes. These distributions as extracted from the CELEX lexical database are based on a count of 42,380,000 Dutch wordforms and 17,979,343 English wordforms. Table 7 shows that in both languages more than half of the wordforms are nouns and around one quarter are verbs. For these categories, Dutch has approximately 4% more wordforms than English. In both languages there are around 12% adjectives and less than 5% adverbs. An important difference is the relative frequency of the ambiguous categories: 94.5% of the wordforms are unambiguous in Dutch and 90% in English. The only ambiguous category in Dutch that outnumbers its English counterpart is the Adj/V category: Dutch has relatively many V/Adj wordforms, mainly participles.

INSERT TABLE 7 ABOUT HERE

Each word was encoded according to the ONC + Stress scheme that appeared as the most powerful in the previous experiments, that is, every item was represented as a syllabified string of phonemes and the stress level of each syllable was also added.

Table 8 displays the success scores for English and Dutch. The global success score is 66.62% for English and 71.02% for Dutch, which is well above chance level in both cases. The unambiguous categories reach an accuracy of 71.77% in English and 74.21% in Dutch. In both languages ambiguous categories are hard to predict: a success score of 20.6% in English and 15.69% in Dutch. For the ambiguous categories, Table 8 contains a second success score between brackets. That score is calculated using the CELEX frequencies for wordforms: if a wordform is ambiguous between two categories, e.g., the wordform could be a noun as well as a verb, the noun and the verb reading are usually not equally frequent. In calculating the bracketed success score, we took this frequency difference into account in this sense that if the algorithm predicted the most frequent category, that prediction was taken to be correct. In doing so, we allowed for underextension in category assignment, a phenomenon documented in children's language by Nelson (1995). When we allow for underextension, the algorithm's success score increases to 68.42 % for English and to 71.78 % for Dutch.

INSERT TABLE 8 ABOUT HERE

The results of this experiment show that the segmental material enriched with information about a wordform's stress pattern allows to predict its word class with an accuracy of up to approximately 70%. In comparison with the previous experiments, in which only the categories noun and verb were involved, this success score is significantly lower (approximately 78% for English and 84% for Dutch). However, a success rate of 7 out of 10 means, at least, that there is a more than arbitrary relationship between phonology and grammatical class, a relationship that the algorithm can exploit given all open class categories in the language.

A consistent finding in the experiments is that in Dutch the link between phonology and grammatical class is significantly more transparent than in English. When we compare the success scores for nouns and verbs (for instance the ONC + Stress enconding, see Table 4 for English and Table 6 for Dutch), Dutch scores consistently higher than English (nouns: 84% for English vs. 85% for Dutch; verbs: 79% for English vs. 84% for Dutch; noun/verb: 26% for English vs. 30% for Dutch; overall success score: 78% for English vs. 83% for Dutch). The results in Table 8 which comprise all open class wordforms point in the same direction: the overall success score for Dutch is significantly higher than the one for English, and the same also holds for the success scores for the individual categories. Adverbs are the main exception to this observation: here English (80%) scores significantly better than Dutch (22%), which is mainly due to the fairly consistent marking of English adverbs with the suffix -ly.

The difference in transparency between English and Dutch was also found in related work (see Gillis, Durieux and Daelemans 1996) where the main focus was on differences in morphological structure and its impact on the phonology/grammatical class connection in English and Dutch. In machine learning experiments using the IBL algorithm, Gillis et al. (1996) investigated the predictability of the grammatical class of morphologically simplex and complex wordforms. More specifically, (a) monomorphemes and morphologically complex words (compounds and derivations) and (b) uninflected and inflected wordforms were compared. A combination of these factors yielded four categories for each language. The results of the machine learning experiments in which an ONC + stress encoding was used with 8,000 training items in each condition, are summarized in Table 9

INSERT TABLE 9 ABOUT HERE

A comparison of the results in Table 9 shows that in each corresponding cell, Dutch scores significantly higher than English. Irrespective of the level of morphological complexity represented in the learning material, the relationship between the phonological structure of wordforms and their grammatical class was easier to detect in Dutch than in English. Hence, the interface appears to be more transparent in Dutch.

In addition it can be seen in Table 9 that in both languages compounding and derivation lead to better results: in English the grammatical class of morphologically simplex words can be accurately predicted in 51% of the cases and the success rate increases to 59% for morphologically complex words. A similar picture holds for Dutch: an increase from 79% for simplex words to 89% for complex words.

Inflection does not have the same effect in both languages. In English it has a disambiguating effect: the success rate of simplex and complex words increases when inflection is added: from 51% to 64% for simplex words and from 58% to 59% for complex words. Dutch shows the reverse picture: inflection appears to add ambiguity, and hence the success rates decrease: from 79% to 74% for simplex words and from 89% to 64% for complex words.

In all the previous experiments two important factors were disregarded, viz. (a) the amount of training data, and (b) the frequency of the wordforms. Indeed, all the experiments were performed with 5,000 training items. It may well be that, with 5,000 items, the algorithm reached its peak performance in the noun-verb experiments in which only two unambiguous and one ambiguous category had to be predicted. However, it is very well possible that, when all open classes are involved, the maximum accuracy may not yet have been reached, e.g. because certain categoriare underrepresented in the training data. In the next section, we report on a learning experiment in which the number of training data is systematically increased, thus allowing a study of the algorithm's learning curve.

A second factor that was neglected in the previous experiments was the frequency of the wordforms. The correlation between irregular forms and their relative frequency is a well known phenomenon: irregular forms (e.g., irregular past tense verbforms) tend to be situated in the higher frequency classes, while the lower the frequency of a wordform, the more likely that it is regular. In the previous experiments, the frequency of the wordforms in the training sets was not controlled for. In the next section, the role of frequency will be further investigated.

5.3. Learning Curves and Issues of Item Frequency

In the previous experiments (2 - 6), all datasets contained the same 5,000 wordforms. This methodology was used in order to reliably compare the impact of different coding schemes. In Experiment 7 we turn to the relationship between the number of items in the training set and the accuracy of the predictions: what is the evolution of the algorithms' predictions vis-à-vis the number of training items? In other words, is there a learning effect when more items are added to the training set, or does the algorithm hit upon the right cues with only a relatively small set? A second issue that will be investigated concerns the relationship between the frequency of words and the accuracy of predictions: not only the number of training items but also their frequency may influence the algorithm's performance. In Experiment 8 the relationship between several frequency regions in the lexicon and the algorithm's performance will be examined.

5.3.1. Experiment 7: Learning Curve

The data for this experiment were once more collected from the CELEX lexical database. Datasets with English and Dutch wordforms were incrementally formed, starting with 500 items. At each step 500 items were added. From 10,000 items onwards 1,000 items were added at each step. Each dataset represented a random stratified selection of wordforms in which the relative frequencies of the various categories remained constant and in agreement with their relative frequencies in the entire CELEX database. Wordforms were coded according to the ONC + Stress scheme, i.e., a wordform was represented as a syllabified string of segments and for each syllable the stress level was added.

The results are displayed in Figure 2 and Figure 3. Figure 2 shows the global success scores for English and Dutch as a function of the number of training items (range 500 - 20,000). Figure 3 shows the results for the individual grammatical categories (nouns, verbs, adverbs, adjectives) and the ambiguous categories.

INSERT FIGURE 2 ABOUT HERE

Figure 2 shows that for English as well as for Dutch there is a significant increase of the success score in predicting open class categories: for English the score increases from 57% with 500 training items to 69% when trained with 20,000 items. The success score stabilizes around 69%. For Dutch there is an increase from 59% with 500 items to 75% with 20,000 items. Even at that last point, there is still a slight increase in the success score. The latter point is clearly shown in Figure 3: while the success scores for Dutch and English nouns, verbs and ambiguous categories have stabilized well before the 20,000 end point (Dutch nouns: 83%, English nouns: 79%, Dutch verbs: 78%, English verbs: 74%, Dutch ambiguous: 15%, English ambiguous: 19%), adjectives and adverbs show a different picture. At least for Dutch, there is still an increase of the success score for adjectives (endpoint 64% for adjectives and 40% for adverbs). English adverbs, on the other hand, have reached their peak success score (82%) already with 2,500 items, and the success score for English adjectives has stabilized at around 57%.

INSERT FIGURE 3 ABOUT HERE

5.3.2. Experiment 8: Wordform frequency

Experiment 8 was designed to assess the influence of a wordform's frequency class on the predictability of its grammatical category. In the previous experiment the relative frequencies of the open classes were kept constant, viz. in each training set they reflected the frequency of each category in the entire database. However, all grammatical categories are not equally distributed over the entire lexicon. This can easily be shown as follows. We divided the open class wordforms of the CELEX lexical database into frequency classes according to the scheme developed by Martin (1983) as specified by Frauenfelder, Baayen, Hellwig and Schreuder (1993: 786). Six frequency classes are identified ranging from very frequent wordforms to extremely rare ones: "very frequent/common" (f >= 1:10,000), "frequent / common" (1:100,000 <= f < 1:10,000), "upper neutral" (1:500,000 <= f < 1:100,000), "lower neutral" (1:1,000,000 <= f < 1:500:000), "rare" (1:10,000,000 <= f < 1:1,000,000) and "extremely rare" (f < 1:10,000,000). Table 10 gives an overview of how the open classes are distributed over these frequency classes. For the sake of convenience the frequency classes were labeled

as “n1” (“very frequent/common”) through “n6” (“extremely rare”), and due to the sparcity of the data the classes with the highest frequency wordforms, viz. n1 and n2, were merged (hence, n1/2).

INSERT TABLE 10 ABOUT HERE

First of all, Table 10 shows a close parallel between English and Dutch in that, in all frequency classes, nouns outnumber all other grammatical categories. Moreover, the most dramatic differences between the frequency classes lies in the percentage of nouns and the percentage of the ambiguous category (in which all categories such as N/V, N/Adj, V/Adv, etc. are collected): the percentage of nouns increases from n1/2 to n6 (English: 37.9% to 65.2%; Dutch: 29.4% to 60.2%), while the percentage of the ambiguous categories decreases from n1/2 to n6 (English: 30.76% to 0%; Dutch: 21.76% to 0.72%). This means that in the high frequency classes there are many more grammatically ambiguous words than in the low frequency classes. And in the low frequency classes there are relatively more nouns than in the high frequency classes.

Given the results of the previous experiments that nouns can be best predicted while ambiguous categories are very hard to predict, this picture leads to the following prediction: success scores in the grammatical class prediction task will increase as frequency decreases, or in other words, the success score will be lower for high frequency items than for low frequency items.

This prediction was tested in an experiment. For each frequency class 5,000 wordforms were randomly selected from the CELEX lexical database. The relative frequencies of the grammatical categories were in agreement with the ones shown in Table 10. The wordforms were coded according to the ONC + Stress coding scheme.

INSERT FIGURE 4 ABOUT HERE

Figure 4 shows the success scores per frequency class. It clearly appears that the success score of the algorithm increases as the frequency of wordforms decreases: for both the English and the Dutch data, the success rate is significantly lower for the most frequent wordforms (n1/2) as compared to the most infrequent wordforms (n6). The

trend is monotonous: each step (from n1/2 to n3 to n4 ...) brings about an increase in the global success score. This means that in both English and Dutch more frequent wordforms appear to have a less transparent mapping between their phonology and their grammatical class.

6. Discussion and conclusion

We set out to investigate to what extent grammatical classes can be predicted from the phonological form of a word. Correlations between a wordform's phonological form and its grammatical class were indicated by amongst others, Kelly (1996), and the potency of phonological bootstrapping as a useful strategy for cracking the grammatical code has been suggested in the language acquisition literature. From the point of view of acquisition, phonological bootstrapping seems to be a helpful strategy in principle. There are some examples scattered throughout the literature of how phonological bootstrapping may work in first language acquisition. A case in point is provided by De Haan, Frijn and De Haan (1995) who show that in disentangling the intricate system of verb-second placement in Dutch, children appear to use the syllable structure of the verb in discovering the relationship between verb placement and the verb's morphology. At some point in the acquisition process children work with the (partially correct) idea that disyllabic trochaic verbs (non-finite verbforms in the adult language) are to be placed in final position and monosyllabic verbs (finite verbs) are to be placed in second position. This strategy is an overgeneralization but it is a good example of how phonology may be helpful in solving a morpho-syntactic problem at a particular point in the acquisition process (see also Wijnen and Verrips 1998).

In this paper, we addressed the question of how far the language learner can get in exploiting phonological bootstrapping as a strategy in acquisition. If phonological bootstrapping is a useful strategy, to what extent can it assist the child in cracking the grammatical code? For this purpose we conducted machine learning experiments that enable us to draw the boundaries of the strategy: if phonological material can be used by the learner to predict the grammatical category (or categories) of the wordforms he encounters, how accurate will these predictions be? In machine learning experiments

specific variables can be systematically varied and the consequences of the variation can be closely monitored, so that the ultimate quantitative consequences of particular hypotheses can be investigated.

The first part of this paper explicitly addressed such a hypothesis, viz. Kelly's (1996) overview of phonological cues that distinguish nouns and verbs in English. The predictive power of those cues was tested in machine learning experiments, in which the learning system had to predict the grammatical class of unseen wordforms based on prior exposure to a representative sample of wordform/category pairs. By varying the number and nature of the phonological cues in the input representation of wordforms, the relative impact of each cue on prediction accuracy could be evaluated.

In a first experiment, the predictive value of stress was investigated, since this feature was claimed to be a reliable indicator of grammatical class. Our results indicated that this claim could only be supported for a very limited subset of the lexicon, viz. for disyllabic wordforms which are orthographically ambiguous between a noun and a verb reading. For larger subsets of the lexicon, the predictive value of stress was shown to be considerably lower. The "less reliable cues", viz. "vowel height", "consonant quality", and "number of segments" were also tested. When considered individually, none of these cues turned out to be good predictors of grammatical class, although vowel height (a cue denoted by Kelly as "weak") was found to yield better predictions than the "strong" cue "stress". A combination of these cues, however, resulted in a significant increase in predictive accuracy, which confirms Kelly's (1996) hypothesis that individually weak cues may put stronger constraints on grammatical class when considered collectively. A similar conclusion was reached by Shi et al. (1998) with respect to the identification of open and closed class words.

These experiments relied on the *a priori* identification of the relevant cues: the learner 'knew' in advance which dimensions were relevant for solving the task of grammatical category assignment, since the cues were precoded in the learning material. When that information was removed from the learning material, a 'raw' phonological encoding appeared: a syllabified string of segments. This 'raw' phonological encoding yielded significantly better results than any of the encodings in terms of more abstract phonological cues. Augmenting this phonological encoding with stress information brought about a significant increase in performance; adding the other

cues had no such effect. These findings are taken to indicate that the cues which had been identified in the literature do not cover all relevant dimensions of the problem domain.

Overall, the results of these experiments strongly support the claim that for English nouns and verbs there is a more than arbitrary relation between phonological form and grammatical category. In almost eight out of ten cases the algorithm can predict the grammatical category of a wordform when the only relevant information is a syllabified string of segments and the stress pattern of the word.

In Experiments 4 and 5 these findings were replicated for Dutch: for Dutch nouns and verbs, a combination of phonological cues yields more accurate predictions than those obtained for each cue individually. And for Dutch too, better performance was obtained with a 'raw' segmental encoding than with an encoding that relies on phonological cues abstracted from the segmental material.

What is the validity of the cues identified by Kelly (1996)? As the comparison of the different encodings showed, valuable information is lost by abstracting away from the segmental details. This could mean that either the segmental representation is the most appropriate one for the task of predicting grammatical category membership, or that the cues identified by Kelly constitute only a subset of the relevant dimensions. In this case, additional cues should be identified in order to obtain an accuracy comparable to that reached with the segmental encoding. That this may prove to be a valuable enterprise is suggested by the encoding of the Dutch material in terms of Kelly's cues. The outcome of Experiment 4, in which the Dutch data were coded in terms of the cues originally designed for English, was at least surprising: the predictions for Dutch nouns and verbs were significantly better than those for their English counterparts. Of course, Dutch and English are typologically very close, and thus far-reaching conclusions about a more than language specific validity of Kelly's cues should be avoided given the present evidence. It is clear that experiments similar to the ones reported on in this paper should be carried out with data from a typologically more varied sample of languages. Nevertheless, the outcome of our experiments shows that any statement about the idiosyncracy of the link between phonology and grammatical class should be pronounced with care.

A second general conclusion that can be drawn from our experiments (especially Experiments 6 and 7) is that the phonology - grammatical class link can be generalized to all open class items in English and Dutch. ForDutch, the prediction of the categories noun, verb, adjective and adverb reached an accuracy of 71% with 5,000 training items and 75% with 20,000 training items. For English, prediction accuracy was 67% with 5,000 items and 69% with 20,000 items. At the same time, the ambiguous categories prove to be very hard to predict and cannot be identified above chance level. For the unambiguous categories, the results are very promising: with a relatively small number of examples, fairly accurate identification of the main parts of speech is possible. Adding in more examples further improves the delineation of the various categories.

A third general conclusion is that accuracy of the prediction of grammatical classes on the basis of phonological characteristics differs from language to language. In all the experiments, the success scores for Dutch are superior to those obtained for English. This means that the mapping from segmental representations to grammatical classes is more transparent in Dutch than it is in English. A counterexample to this general trend is the class of English adverbs: the success score for English adverbs is above 80% early in the learning curve. The well-known fact that a high number of English adverbs end in -ly, which is reflected as such in the segmental encoding of the learning material, provides a very straightforward mapping between phonology and part-of-speech, a transparent mapping that is lacking for Dutch adverbs.

The conclusion of our machine learning experiments is that ‘in principle’ the link between phonology and grammatical class can be exploited with a certain degree of confidence. The question now is whether and how children exploit this link. Our experiments cannot provide an adequate answer to this question, because in the present set of experiments a number of abstractions were made that prevent us from going beyond the ‘in principle’ answer.

The approach taken in this study carries a number of abstractions. First of all, the machine learning algorithm used in the experiments incorporates a supervised learning technique. The learner is first trained with precategorized material and is then tested on its ability to generalize from that material. Of course, children do not get similar input: they not receive a properly tagged corpus from which they can generalize.

They do not get a part-of-speech label with every single word they hear. A non-supervised approach will have to be taken in order to relieve this constraint (see, e.g., Shi et al. 1998).

A second abstraction is that the phonological information is presented in isolation. The learner only 'sees' or 'hears' a wordform's phonological form and does not have access to other information, such as the syntactic patterns in which it occurs, or its meaning. It may well be the case that a child uses all these knowledge sources concurrently as a bootstrap: as De Haan et al. (1995) and Krikhaar and Wijnen (1995) have shown that children are able to use correlated syntactic, semantic and phonological cues in solving particular grammatical problems.

Taken together, these two abstractions may lead us in the right direction for future experiments in a supervised framework. Indeed, children do not get a part of speech label with each word they hear, but at least for some words they get appropriate contextual information to make sense of the type of entity involved (i.e., semantic information). At the same time they get structural information in the sentences they hear: those sentences carry distributional information of target words. These phonological, structural and semantic cues may not be reliable cues in isolation, but taken together, they may lead the learner towards a reliable delineation of major word classes. This, of course, remains a research direction that needs to be explored further.

The experiments reported in this paper raise still other important questions that were sharply identified and illustrated with quantitative data. First of all, the algorithm is tested with several thousands of wordforms. Does this mean that phonological bootstrapping only becomes a useful strategy when a child 'knows' (presumably, comprehends) a few thousands of words? Given the evidence on bootstrapping as suggested in the studies of De Haan et al. (1995) and Krikhaar and Wijnen (1995) this does not appear to be the case: even with a relatively modest lexicon, two-year-olds appear to rely on phonological characteristics of words to resolve structural problems. However, they do not appear to display an across the board strategy (e.g., they do not use phonology to solve grammatical class membership, as illustrated in this paper). Instead, their use of phonological cues appears to be geared towards solving particular structural problems (e.g., they use phonology in a task oriented manner, such as relating

the phonological characteristics of particular words to solve a specific structural puzzle, as illustrated by De Haan et al. (1995) and Krikhaar and Wijnen (1995)).

Another problem raised by our experiments concerns the role of frequency. In Experiment 8 the role of token frequency was investigated. The main result was that the more frequent a wordform is the less transparent the mapping between its phonology and grammatical category is. For phonological bootstrapping in language acquisition this means that infrequent words are better for discovering the intricacies of the phonology - syntax interface than high frequency words. At first sight this poses a serious problem, since it is at least intuitively clear that the words children hear stem from the high frequency regions of the lexicon. It is exactly in the high frequency regions that the most ambiguous wordforms appear (see Table 11): up to 30% in English and more than 20% in Dutch, and IBL's performance was poor on ambiguous wordforms. At second glance, the algorithm opted for a strategy of predicting unambiguous categories instead of ambiguous ones. Moreover, IBL predicted the more frequent category that agreed with the phonological form of the word. And this also seems to be the strategy adopted by children: children appear to bypass the problem of ambiguous wordforms in the sense that they employ a 'one form - one function' strategy according to Nelson (1995), who specifically addresses the issue of dual category forms: "For most forms, most children seemed to obey the one form - one function principle in production, ..." (Nelson 1996: 246). This result is promising but it leaves the question unanswered of how ambiguous categories are eventually acquired.

The outcome of the experiments reported in this paper should thus be seen as an indication of how far the principle of phonological bootstrapping can lead the child in discovering grammatical categories, which is only a first step in showing how phonological information is used together with distributional patterns, semantic information, and other kinds of information.

References

Aha, D., D. Kibler and M. Albert. 1991. "Instance-based learning algorithms". *Machine Learning* 6: 37-66.

Baayen, H., R. Piepenbrock & H. van Rijn. 1993. *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA.: Linguistic Data Consortium.

Bates, E. and B. MacWhinney. 1989. "Functionalism and the competition model". In B. MacWhinney and E. Bates (eds.), *The Crosslinguistic Study of Language Processing*. New York: Cambridge University Press, 3-73.

Cartwright, T. and M. Brent. 1996. "Early acquisition of syntactic categories". Ms.

Daelemans, W. and A. van den Bosch. 1992. "Generalisation performance of backpropagation learning on a syllabification task." In M. Drossaers and A. Nijholt (eds.), *TWLT3: Connectionism and Natural Language Processing*. Enschede: Twente University, 27-38.

Daelemans, W., S. Gillis & G. Durieux. 1994. "The acquisition of stress: a data-oriented approach". *Computational Linguistics* 20: 421-451.

De Haan, G., J. Frijn and A. De Haan. 1995. "Syllabestructuur en werkwoordsverwerving". *TABU* 25: 148-152.

Finch, S. and N. Chater. 1992. "Bootstrapping syntactic categories". In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society of America*, 820-825.

Francis, W. & H. Kucera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton-Mifflin.

Frauenfelder, U., H. Baayen, F. Hellwig, & R. Schreuder. 1993. "Neighborhood density and frequency across languages and modalities". *Journal of Memory and Language* 32: 781-804.

Gentner, D. 1982. "Why nouns are learned before verbs: linguistic relativity versus natural partitioning". In S. Kuczaj (ed.), *Language Development: Language, Culture, and Cognition*. Hillsdale: Erlbaum, 301-334.

Gillis, S. and G. Durieux. 1997. "Learning grammatical classes from phonological cues. In A. Sorace, C. Heycock and R. Shillcock (eds.), *Language Acquisition: Knowledge Representation and Processing*. Edinburgh: HCRC.

Gillis, S., G. Durieux and W. Daelemans. 1996. "On phonological bootstrapping: 'l arbitraire du signe revisited". Paper presented at the VIIth International Congress for the Study of Child Language, Istanbul.

Gillis, S., W. Daelemans and G. Durieux. In press. “Lazy Learning”. In J. Murre and P. Broeder (eds.), *Cognitive Models of Language Acquisition*. Cambridge, Mass.: MIT Press.

Gleitman, L., H. Gleitman, B. Landau and E. Wanner. 1988. “Where learning begins”. In F. Newmeyer (ed.), *The Cambridge Linguistic Survey.* New York: Cambridge University Press, Vol. 3, 150-193.

Kager, R. 1989. *A Metrical Theory of Stress and Destressing in English and Dutch*. Dordrecht: Foris.

Kelly, M. 1992. “Using sound to solve syntactic problems”. *Psychological Review* 99: 349-364.

Kelly, M. 1996. “The role of phonology in grammatical category assignment”. In J. Morgan and K. Demuth (eds.), *From Signal to Syntax*. Hillsdale: Erlbaum, 249-262.

Krikhaar, E. and F. Wijnen. 1995. “Children’s categorization of novel verbs: syntactic cues and semantic bias”. In E. Clark (ed.),

Maratsos, M. and M. A. Chalkley. 1980. “The internal language of children’s syntax”. In K. Nelson (ed.), *Children’s Language*. New York: Gardner Press, Vol. 2, 127-214.

Martin, W. 1983. “On the construction of a basic vocabulary”. In S. Burton and D. Short (eds.), *Proceedings of the 6th International Conference on Computers and the Humanities*, 410-414.

Mintz, T., E. Newport & T. Bever. 1995. “Distributional regularities in speech to young children”. In *Proceedings of NELS 25*, 43-54.

Morgan, J., R. Shi & P. Allopena. 1996. “Perceptual bases of rudimentary grammatical categories”. J. Morgan & K. Demuth (eds.), *From Signal to Syntax*. Hillsdale: Erlbaum, 263-283.

Nelson, K. 1995. “The dual category problem in the acquisition of action words”. In M. Tomasello & W. Merriman (eds.), *Beyond Names for Things*. Hillsdale: Erlbaum. 223-249.

Pinker, S. 1987. “The bootstrapping problem in language acquisition”. In B. MacWhinney (ed.), *Mechanisms of Language Acquisition*. Hillsdale: Erlbaum, 399-441.

Quinlan, J. R. 1986. “Induction of decision trees”. *Machine Learning* 1: 81-106.

Shi, R., J. Morgan and P. Allopenna. 1998. “Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective”. *Journal of Child Language* 25: 169-201.

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

Wijnen, F. and M. Verrips. 1998. “The acquisition of Dutch syntax”. In S. Gillis and A. De Houwer (eds.), *The Acquisition of Dutch*. Amsterdam: Benjamins. 223-299.

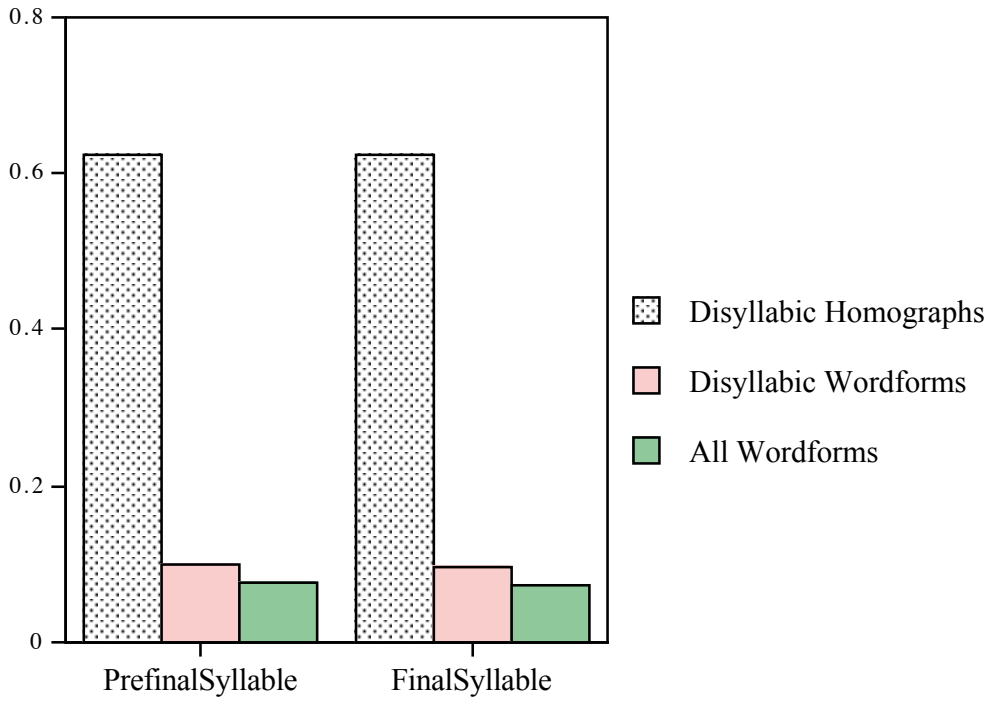


Figure 1: Information Gain values for Disyllabic Homographs, Disyllabic Wordforms and All Wordforms.

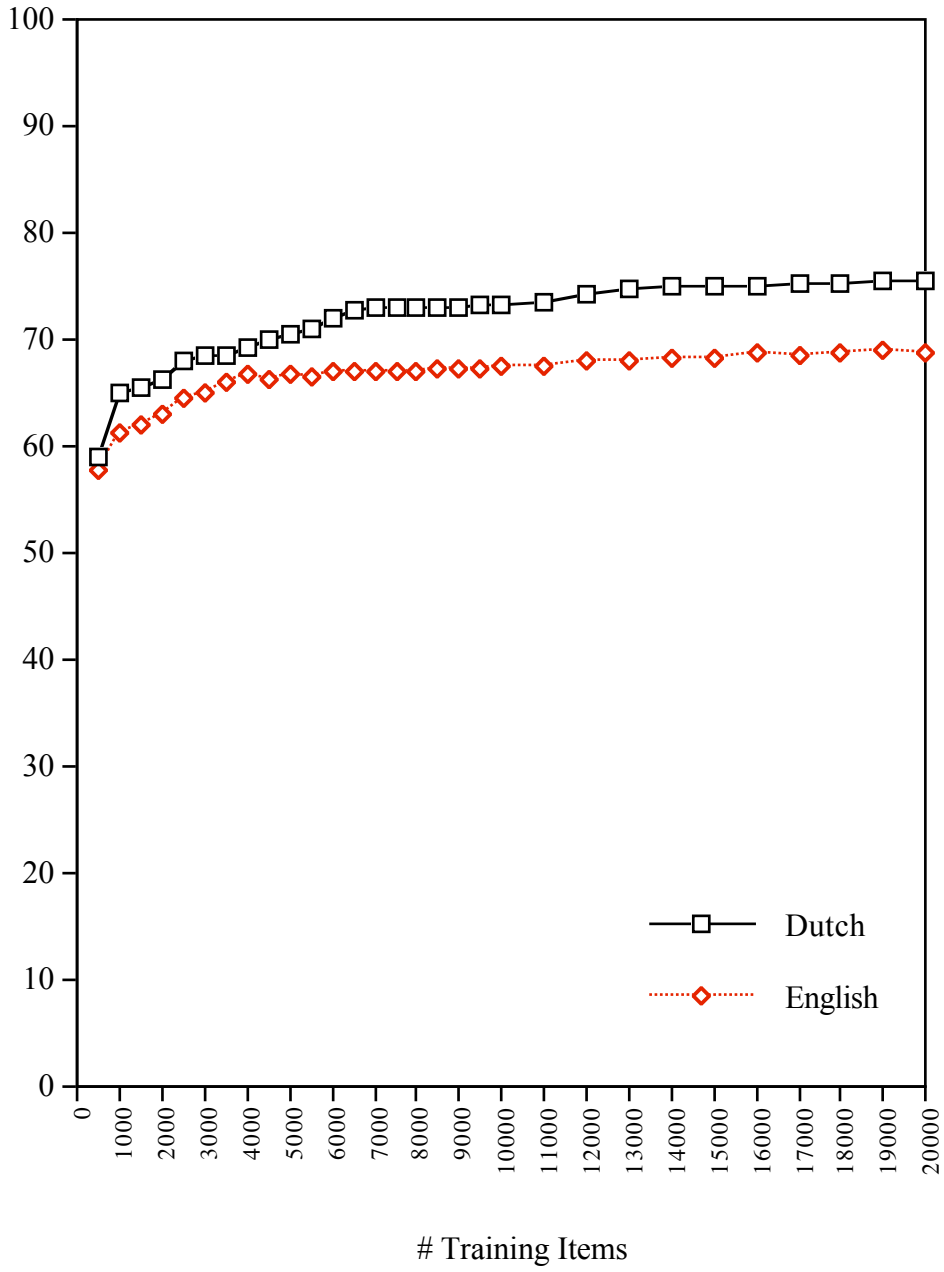
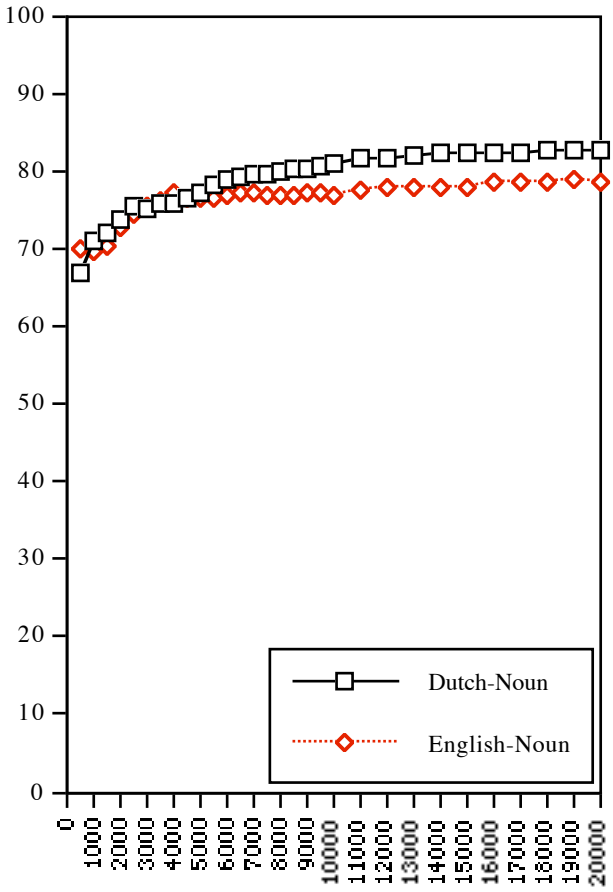
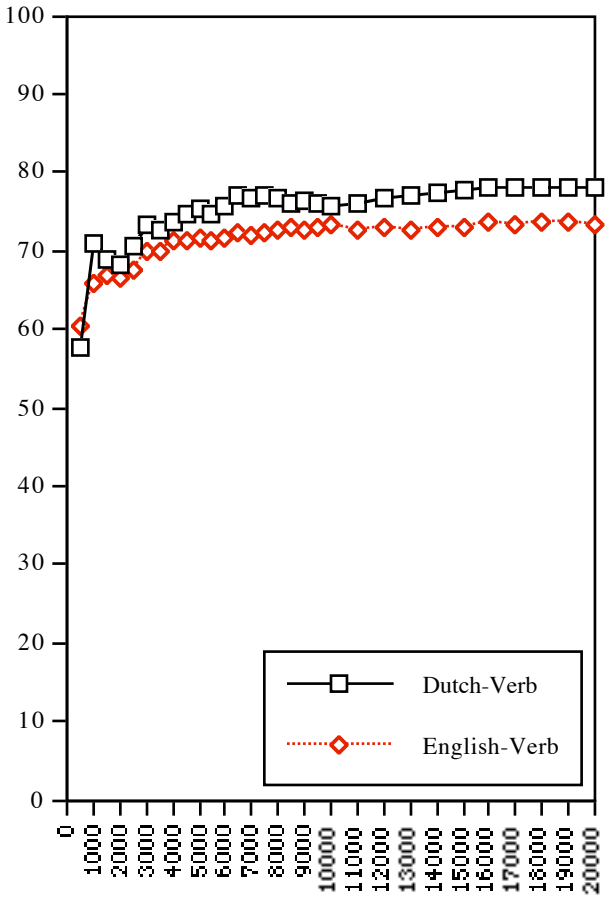
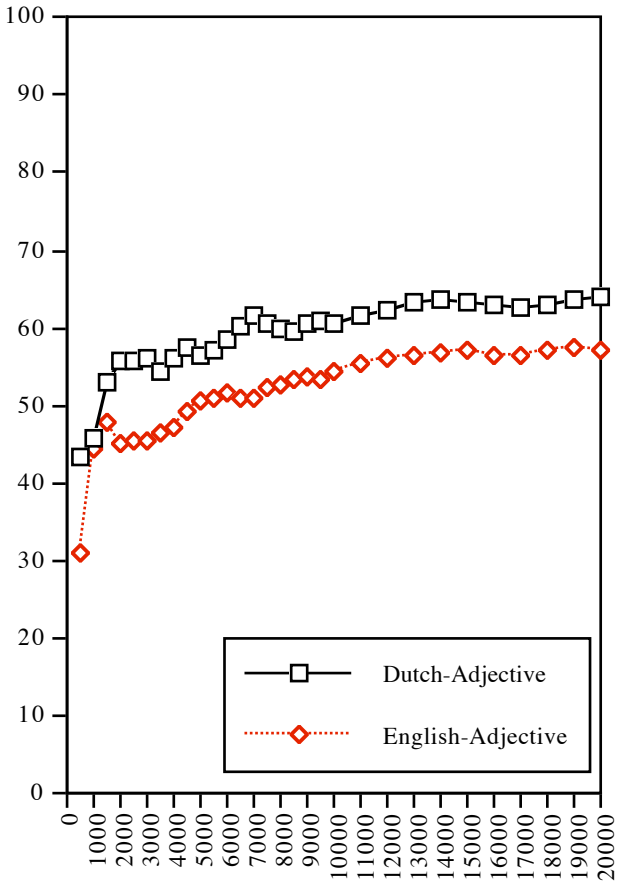
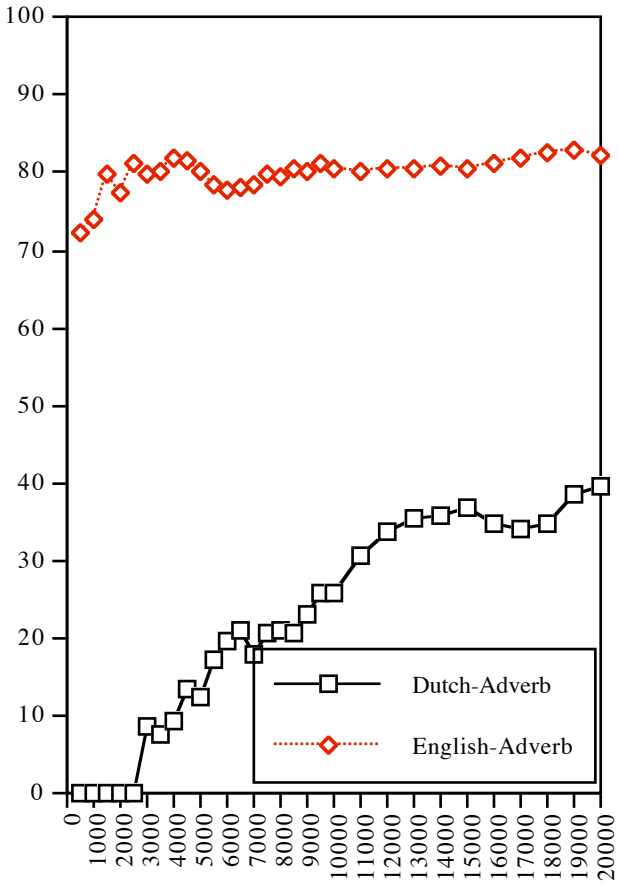


Figure 2: Global success scores for English and Dutch open class wordforms


100
90
80
70
60
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
Dutch-Noun
English-Noun
Training Items


100
90
80
70
60
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
Dutch-Verb
English-Verb
Training Items


100
90
80
70
60
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
Dutch-Adjective
English-Adjective
Training Items


100
90
80
70
60
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
Dutch-Adverb
English-Adverb
Training Items

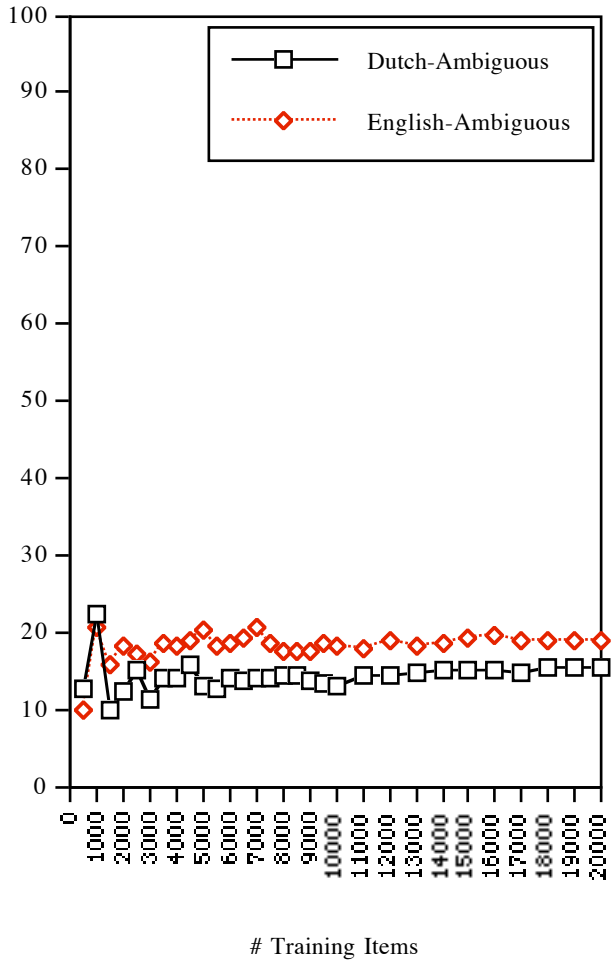


Figure 3: Success scores for English and Dutch open class wordforms

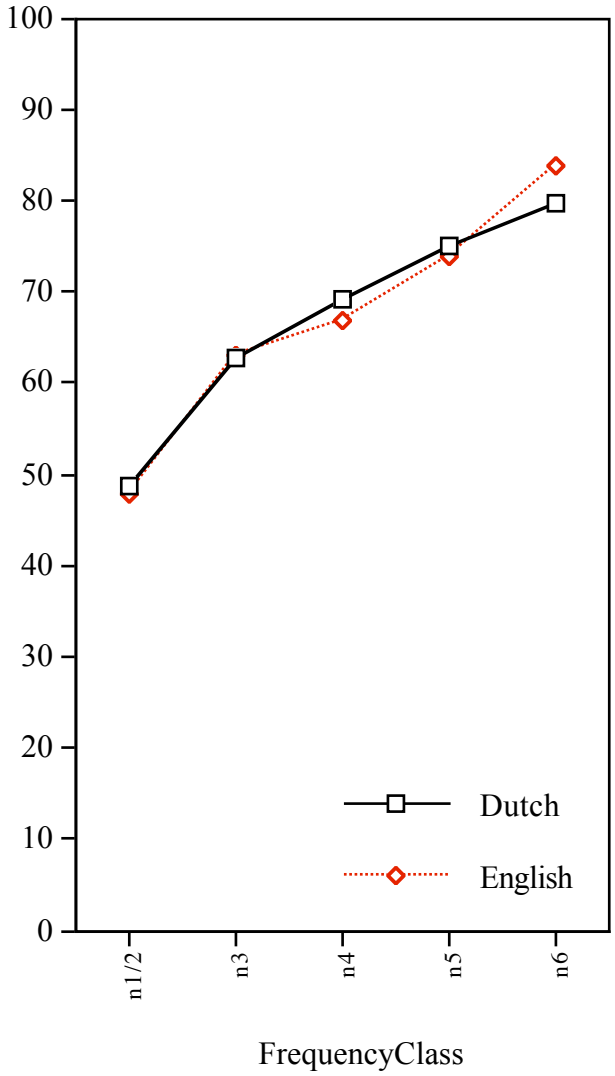


Figure 4: Global Success Scores for Dutch and English Wordforms Relative to the Frequency Classes.

Table 1: Success scores for Disyllabic Homographs

Category	# Wordforms	# Correct	% Correct
N	212	179	84.43%
V	215	182	84.65%
NV	16	5	31.25%
Total	443	366	82.62%

Table 2: Comparison of success scores for Disyllabic Homographs, Disyllabic Wordforms and All Wordforms

Category	Disyllabic Homographs	Disyllabic Wordforms	All Wordforms
N	84.43%	92.84%	94.77%
V	84.65%	38.91%	23.13%
NV	31.25%	0%	0%
Total	82.62%	69.74%	66.18%

Table 3: Success scores for individual cues and a combination of all the cues

Category	Stress	Vowel Height	Consonant Quality	Number of Segments	Combined Cues
N	94.77%	91.61%	87.66%	94.64%	77.59%
V	23.13%	34.04%	33.91%	10.50%	61.37%
NV	0%	0%	0%	0%	11.96%
Total	66.18%	67.38%	64.86%	62.42%	67.68%

Table 4: Success scores for phonological encodings

Category	ONC	ONC + Stress	ONC + Combined Cues	Combined Cues
N	79.24	84.18	83.74	77.59%
V	75.70	79.41	80.65	61.37%
NV	25.80	25.80	26.29	11.96%
Total	73.86	78.04	78.16	67.68%

Table 5: Success scores for individual cues and a combination of all the cues (Dutch data)

Category	Stress	Vowel Height	Consonant Quality	Number of Segments	Combined Cues
N	76.88%	75.67%	85.19%	74.92%	81.77%
V	23.70%	52.29%	11.58%	56.21%	67.97%
NV	25.78%	18.75%	17.97%	0.0%	4.69%
Total	57.94%	66.46%	59.06%	66.80%	75.22%

Table 6: Success scores for Dutch wordforms: phonemic encoding of nouns and verbs

Category	ONC	ONC + Stress	ONC + Combined Cues	Combined Cues
N	83.82%	85.28%	85.97%	77.59%
V	83.66%	84.08%	84.26%	61.37%
NV	29.69%	29.69%	29.69%	11.96%
Total	82.38%	83.46%	83.96%	67.68%

Table 7: Frequency distribution of English and Dutch Wordforms in CELEX

Category	English	Dutch
N	50.46 %	54.4 %
V	23.14 %	27.28 %
Adj	12.14 %	11.92 %
Adv	4.22 %	0.92 %
N/Adj	1.66 %	0.84 %
N/V	6.44 %	2.08 %
N/Adv	0.06 %	0.04 %
Adj/V	0.98 %	2.28 %
Adj/Adv	0.42 %	0.02 %
Adv/V	0.04 %	0 %
N/V/Adj	0.22 %	0.22 %
N/Adj/Adv	0.14 %	0 %
V/Adj/Adv	0.06 %	0 %
N/V/Adv	0 %	0 %
N/V/Adj/Adv	0.02 %	0 %

Table 8: Success scores for all open classes (Dutch and English data)

Category	English	Dutch
N	76.30%	78.60%
V	71.48%	75.37%
Adj	50.58%	55.70%
Adv	80.09%	21.74%
N/Adj	1.20% (40.96%)	4.76% (45.24%)
N/V	29.50% (42.24%)	17.31% (35.58%)
N/Adv	0.00% (0.00%)	0.00% (50.00%)
Adj/V	8.16% (28.57%)	20.18% (20.18%)
Adj/Adv	14.29% (28.57%)	0.00% (0.00%)
Adv/V	0% (0%)	/ [a]
N/V/Adj	0.00% (27.27%)	0.00% (9.09%)
N/Adj/Adv	0.00% (0.00%)	/
V/Adj/Adv	0.00% (0.00%)	/
N/V/Adv	/	/
N/V/Adj/Adv	0.00% (0.00%)	/
Total	66.62% (68.42%)	71.02% (71.78%)

[a] A slash means that this category was not represented in the data.

Table 9: Success score for wordforms of different morphological complexity in English and Dutch

English		
	Uninflected	Inflected
Morphologically Simplex: Monomorphemes	51%	64%
Morphologically Simplex & Complex: Monomorphemes & Compounds & Derivations	58%	59%
Dutch		
	Uninflected	Inflected
Morphologically Simplex: Monomorphemes	79%	74%
Morphologically Simplex & Complex: Monomorphemes & Compounds & Derivations	89%	68%

Table 10: Distribution of grammatical classes over frequency classes

English		Frequency Class				
		n1/2	n3	n4	n5	n6
	%Nouns	37.9	43	46.48	54.28	65.2
	%Verbs	17.64	23.9	25.3	24.72	17.26
	%Adjectives	8.64	11.6	14.52	13.02	12.18
	%Adverbs	5.06	3.68	3.4	3.68	5.36
	%Ambiguous	30.76	17.82	10.3	4.3	0
Dutch		Frequency Class				
		n1/2	n3	n4	n5	n6
	%Nouns	39.4	45.4	49.4	58.52	60.2
	%Verbs	21	25.72	27.56	25.16	26.84
	%Adjectives	13.2	14.96	15.24	12.7	11.9
	%Adverbs	4.64	1.58	0.94	0.54	0.34
	%Ambiguous	21.76	12.34	6.86	3.08	0.72