Predicting Grammatical Categories[*]

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

Abstract

We investigate to what extent the grammatical class(es) of a word can be predicted exclusively on the basis of phonological and prosodic information. We report on several experiments with an artificial learning system which has to assign English nouns and verbs to their appropriate grammatical class using various types of phonological and prosodic information. We replicate the claim articulated by Kelly (1996) that specific phonological cues are particularly good predictors of grammatical class. Our results indicate that these cues are partially predictive for the problem at hand. We show experimentally that 'raw' segmental and suprasegmental information (word stress) convey most information about the 'phonology - syntax' interface.

[*] Preparation of this paper was supported by a VNC project of NFWO - NWO (contract number G.2201.96), and by a GOA grant.

Predicting Grammatical Categories

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

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: under this hypothesis, 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, 1984; 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 & Chalkey, 1980; Finch & Chater, 1992; Mintz, Newport & Bever, 1995). 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, e.g. "open" vs. "closed" class words or 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). Another common characteristic is that, typically, it is argued that there is only a partial, i.e. 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. On the other hand, syntactic criteria remain the ultimate determinants of grammatical classes while correlating phonological information could, at best, be supportive, when both agree, or downright misleading, when they do not. More fundamentally, the old Saussurean adage of the “arbitrariness of the sign” casts doubts upon the very existence of phonological cues to grammatical classes. Summarizing, 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 (a) phonological correlates to grammatical classes do exist for English, and that (b) people are sensitive to these correlates, even if they are only weakly diagnostic of grammatical classes. A fairly reliable correlate is the stress pattern of disyllabic words: an examination of 3000 disyllabic nouns and 1000 disyllabic verbs drawn from Francis & Kucera (1982), revealed that 94% of nouns have a trochaic (initial) stress pattern , whereas 69% of verbs display a 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 had either to construct sentences with a stressed disyllabic word, or read target sentences containing a disyllabic non-word in either nominal or verbal position, showed an outspoken preference to link iambic words with the category *verb* and trochaic words with the category *noun*.

Another cue mentioned by Kelly (1996) is the number of a word’s syllables: 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 duration, vowel quality, consonant quality and phoneme number (Kelly, 1996; p. 252): (a) nouns are generally longer than verbs (controlling for syllable number), (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. For discussion of the specific machine learning algorithm used, we refer the reader to Section 3. Our first objective will be to asses the predictive value of stress; in order to do this, we will run a series of experiments covering increasingly larger segments of the 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 words containing a different number of 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 the lexical space. A second objective will be to assess the value of the "less reliable cues". To this end, we will set up a test for each of the observed minor cues (with the exception of duration), 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, finally, is to determine whether the phonological makeup of words restricts grammatical class when no a priori identification of relevant cue(s) is performed, and only the raw phonological material is used. These experiments will be discussed in Section 4.

3. The Learning Algorithm

In this study, we used a modified version of Instance-Based Learning (IBL, Aha et al. 1991). This algorithm falls within the class of supervised learning algorithms: the system is trained by presenting a number of input patterns (i.e. wordforms, coded as a vector of features) together with their correct classification (i.e. their appropriate grammatical class). Testing the system consists in presenting previously unseen wordforms, suitably encoded as feature vectors, and have the system predict their grammatical class. A distinguishing feature of this *lazy learning* algorithm is that no explicit abstractions, such as rules or decision trees, are constructed during training. Instead, a selection of the examples encountered during training is retained, and these examples themselves are used to classify new inputs.

The learning component of IBL is thus set up as follows: during training, pre-categorized items are presented in an incremental way to the learning component. If the item was not already encountered earlier, a new memory record is created in which the item (a word) and its proper categorization (its grammatical class) are stored.

During the test phase, the performance component carries out a required task. In this case, IBL has to predict the grammatical class of a novel word, i.e. a word not encountered during training. For this the system relies on an explicit procedure for determining the similarity of a test item with the items present in memory. If verifying memory does not yield an exact match, the similarity of the test item with all items kept in memory is computed and a category is assigned based on the category of the most similar 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. This metric treats all features as equally important, though. We extended the algorithm with a technique for automatically assigning a degree of relative importance of features. The concept of Information Gain (see e.g., Quinlan 1986) was used for this aim. 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. This aim is reached by incorporating the information gain of each feature as a weight in the similarity metric. (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. 1997).

4. Experiments

4.1 Data and Method

All data for the experiments were taken from the CELEX lexical database (Baayen, 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 80,531 wordforms, belonging to 29,967 lemmas. For the experiments, we restrict the database to noun and verb lemmas encountered at least once in the Collins/Cobuild corpus, which yields 21,571 noun lemmas and 5,257 verb lemmas. Those lemmas account for 30,122 noun wordforms, 15,446 verb wordforms and 5,238 wordforms which are phonologically ambiguous between noun and verb. This restricted database of nouns and verbs will be called 'the database' in what follows.

All experiments were run using the 'leaving-one-out' method (Weiss & Kullikowski 1991) to get the best estimation of the true error rate of the classifier. 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. 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 investigated 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 was 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, were coded as “N” for *noun*, “V” for *verb*, and “N/V” for ambiguous wordforms.

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, it does not imply that the prediction is perfect. 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.

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%

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 that are othographically 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 (i) a random selection of all disyllabic words, and (ii) a random selection of all wordforms of the database.

(i) 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 5000 items was made, consisting of 2933 nouns, 1545 verbs and 522 ambiguous wordforms.

(ii) The dataset was expanded to include wordforms of up to four syllables (henceforth: "All Wordforms"): from the database a random stratified selection of 5000 items was made, containing 2906 nouns, 1594 verbs and 500 ambiguous wordforms. 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 less 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.

Table 2 displays the results of the learning experiment with these two new datasets. In comparison with the Disyllabic Homographs, the overall success scores for the Disyllabic Wordforms and All Wordforms are far inferior: 59% and 55% resp. This drop of accuracy is most spectacular for verbs: whereas in the Disyllabic Homographs dataset verbs were correctly classified in almost 85%, this level of accuracy drops to 50% (Disyllabic wordforms) or less (All Wordforms).

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

Category	Disyllabic Homographs	Disyllabic Wordforms	All Wordforms

N	84.43%	71.26%	67.65%
V	84.65%	50.94%	42.66%
NV	31.25%	19.16%	25.40%
Total	82.62%	59.54%	55.46%

Taken together, these results show that stress is a good predictor in the case of disyllabic homographs, but already a far less reliable 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 claim 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 is rather dubious.

4.3. Experiment 2: Less reliable cues

Apart from stress, a number of "less reliable cues" were identified (see section 2): (a) nouns 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 these 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 provide actual data of acoustic measurements to allow a suitable encoding.

For the sake of comparison we use the same random stratified sample of 5000 words in all the experiments we report in this section, exactly the same dataset that was also used in the experiments with ‘All Wordforms’ in previous section. The sample contains 2906 nouns, 1594 verbs and 500 ambiguous wordforms. Word length varies from one to four syllables. We first describe the actual encoding of the various ‘less reliable cues’ and then we will present a global overview of the results and a detailed comparison of the success scores.

(i) For the first experiment, wordforms were coded for vowel height: for each syllable, a single feature was used, corresponding to the syllable nucleus. Values for this feature were "high" (for the vowels /I, i:, U, u:/), "mid" (for the vowels / E, 3:/) and "low" (for the vowels /{, O:, Q, A:, V/). For diphthongs, a difference was 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 /eI, aI, OI, @U, aU/, and “centering” diphthongs /I@, E@, U@/. Words containing less than four syllables were padded to the left with null features. Finally, the number of syllables was added as a fifth feature.

(ii) For the second experiment, wordforms were coded for consonant quality. Here, two features per syllable were 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 were left-padded with null features, and the number of syllables was added as an additional feature, yielding nine features in all.

(iii) For the third experiment, five features were used, indicating the number of segments per syllable and the number of syllables.

(iv) 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 “stress” is mentioned (see Table 2 in previous section), followed by “vowel height”, “consonant quality”, and “number of segments”. The last column contains the success scores for the combination of these cues.

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

Category	Stress	Vowel Height	Consonant Quality	Number of Segments	Combined Cues
N	67.65%	73.30%	69.03%	66.31%	76.26%
V	42.66%	52.95%	44.48%	41.91%	63.17%
NV	25.40%	27.60%	25.80%	24.20%	27.60%
Total	55.46%	62.24%	56.88%	54.32%	67.22%

An analysis of the global success scores (‘Total’ in Table 3) reveals that taken individually, “vowel height” is the best predictor of grammatical class. The success score for this cue is significantly higher than the success score for the other cues (as measured with McNemar’s Symmetry Chi Square Test). The global success scores for the other three cues do not differ significantly, i.e. predicting a word’s grammatical class on the basis of its stress pattern, the quality of the consonants or its number of segments per syllable yields comparable results. 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. It appears quite clearly that “stress” has a comparable predictive power to “consonant quality” and “number of segments”. “Stress” is a significantly less powerful predictor than “vowel height”.

A second striking 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, the same rank order appears: “vowel height” is the best predictor for the three categories, followed by “consonant quality”, “stress” and “number of segments” in that order.

A third striking fact is that, as hypothesized by Kelly (1996), a combination of the cues significantly better predicts the grammatical class of a word than any singly 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 67%, which is significantly better than the success scores of the individual cues (as measured with McNemar’s Symmetry Chi Square Test).

Turning to the individual categories, it appears that the phonological cues used in this experiment are better predictors of English nouns than verbs or the ambiguous noun/verb 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 76% of the nouns were correctly classified while 27% of the ambiguous words were identified as such. The highest success score for verbs is 63%; verbs gain most from a combination of the individual cues: the best score for the individual cues is almost 53% for “vowel height”, which is significantly less than the 63% for the combined cues. The ambiguous noun/verb category is the hardest to predict on the basis of the cues selected.

However, for most wordforms in this category, noun or verb usage are not equally likely. If we relax the criteria for correct prediction and consider a prediction of “N” (or “V”) correct if “N” (or “V”) is the more frequent usage for that wordform, a different picture emerges: under these criteria, prediction accuracy for ambiguous wordforms is on a par with that for verbs using the “combined cues” encoding.

Table 4: Success scores for 'less reliable cues' allowing for underextensions.

Category	Vowel Height	Consonant Quality	Number of Segments	Combined Cues
N	73.30%	69.03%	66.31%	76.26%
V	52.95%	44.48%	41.91%	63.17%
NV	62.00%	63.40%	61.00%	62.70%
Total	65.68%	60.64%	58.00%	70.72%

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 would only make an ‘educated guess’, i.e., such as predicting the most frequent category, a success rate of 58% was to be expected. The mere fact that IBL reaches a score of 67% is suggestive for 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. This issue will be taken up in the next experiment.

4.4 Experiment 3: Phonological encoding

In Experiment 2, all encodings were obtained by extracting 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” indicated that considering more than one dimension at the same time gave rise to more accurate predictions. To investigate the question to what extent relevant dimensions are picked up by the learner without a priori identification, we set up an experiment in which the 'raw' phonological material was used. In order to asses the importance of the cues identified by Kelly (1996) - see Experiment 2 - those cues were used to 'enrich' the phonemic representation. If the 'a priori' cues define the only relevant dimensions, we expect equal or poorer performance from the phonological encoding. Equal performance would indicate that the phonological 'a priori' cues provide all the necessary information for category assignment to which the segmental material does add any relevant information. Poorer performance may occur because the relevant oppositions defined by the cues are buried under "feature noise". If, on the other hand, the learning algorithm is capable of capitalizing on other information in the phonological encoding, the impact on the results may go in either direction.

As in Experiment 2, the same random stratified data set of 5000 items was used to facilitate comparison of results. In all three encodings, words were padded to the left. The dataset was coded as follows:

(i) For the first encoding, three features per syllable were used, corresponding to the onset, the nucleus and the coda (henceforth: "ONC"). Number of syllables was added as an additional feature.

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

(iii) 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 5. For the sake of convenience, we add the success scores of the Combined Cues from Experiment 2 (see Table 3). A first interesting finding comes from a comparison of the success scores of the ONC and the ONC + Combined Cues encodings. It is quite clear that the global success scores (73.9% vs. 67.2%) as well as the success scores for the individual categories (N: 80.6% vs. 76.26%; V: 73.8% vs. 63.2%; NV: 35.6% vs. 27.6%) are significantly higher (as measured with McNemar's Symmetry Chi Square Test) when the learning material is presented as a syllabified string of segments. In other words, 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 similar cues as those established by Kelly and/or if the algorithm bases its predictions on (entirely) different information.

A second relevant finding is that stress is indeed a relevant factor for predicting nouns and verbs: the success score for ONC, viz. 73.9%, increases to 76% when in addition to the segmental material also the suprasegmental information is added in the encoding of the training material. On the other hand, adding the other cues, viz. "vowel height", "consonant quality" and "number of segments" does not bring about a significant increase in accuracy. This latter finding seems to indicate that those cues do not add any relevant information that IBL can use to base its predictions on, or in other words, the higher level information that these cues bring to the task of category assignment is already 'used' by IBL on the basis of the segmental material.

Table 5: Success scores for phonological encodings

Category	ONC	ONC + Stress	ONC + Combined Cues	Combined Cues
N	80.56%	82.52%	82.11%	76.26%
V	73.78%	76.85%	77.67%	63.17%
NV	35.60%	35.40%	37.00%	27.60%
Total	73.90%	76.00%	76.18%	67.22%

In conclusion, the experiments reported 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.c. syllabified strings of segments and the stress pattern of the wordform. 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.

5. Conclusions

We set out to investigate to what extent the major grammatical classes 'noun' and 'verb' can be predicted from the phonological form of a word. Correlations between a wordform’s phonological form and its grammatical class were indicated by several authors and the potency of phonological bootstrapping as a useful strategy for cracking the grammatical code was suggested in the language acquisition literature.

In this study we investigated to what extent the phonological cues established by Kelly (1996) are indeed good cues for grammatical class membership in English. These questions were approached through a number of 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 the 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, i.c. 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.

In a second experiment, a number of “less reliable cues” were examined: “vowel height”, “consonant quality”, and “number of segments”. When considered individually, none of these cues turned out to be good predictors of grammatical class, although vowel height was found to yield better predictions than stress. 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.

In a third experiment, no a priori identification of relevant cues was performed; instead, the syllabified phonological string was used directly. This ‘raw’ phonological encoding proved to yield significantly better results than any of encodings from the previous experiments. Augmenting this phonological encoding with stress information brought about a significant increase in performance; adding in the other cues had no such effect. These findings were 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 our experiments strongly support the claim that, for English, there is a more than arbitrary relation between phonological form and grammatical category. This immediately raises the question if similar correspondences between the segmental and suprasegmental characteritics of words and their grammatical class can be found in other languages than English. Gillis and Durieux (in press) report preliminary results which indicate that in Dutch there are even stronger ties between phonology and grammatical class. If the cross-linguistic validity of this correspondence can be established, the question turns up what the psycholinguistic implications are: what do correspondences of the type discussed here imply for the organization of the mental lexicon and for linguistic processing in general? For instance, one may wonder if it is at all a coincidence that in malapropisms (as discussed by i.a. Fay & Cutler 1977) the target and the error are not only of the same grammatical class, but also share key phonological properties such as stress pattern, number of syllables, etc.? Furthermore, what psycholinguistic implications do these findings have for the psycholinguistic 'reality' of memory-based models of the language learner and the language user? At present, these questions point at promising avenues for future research.

References

Aha, David H., Dennis Kibler & Mark Albert . 1991. Instance-based learning algorithms. Machine Learning 6:37-66.

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

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

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

Daelemans, Walter, Steven Gillis & Gert Durieux. 1994. The acquisition of stress: a data-oriented approach. Computational Linguistics 20:421-451.

Daelemans, Walter & Antal van den Bosch. 1992. Generalisation performance of backpropagation learning on a syllabification task.
Mark Drossaers & Anton Nijholt (eds.), TWLT3: Connectionism and Natural Language Processing, 27-38. Enschede: Twente University.

Fay, David & Anne Cutler. 1977. Malapropisms and the structure of the mental lexicon. Linguistic Inquiry 8: 505-520.

Finch, Steven & Nick Chater. 1992. Bootstrapping syntactic categories. Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society of America, 820-825.

Gillis, Steven, Walter Daelemans & Gert Durieux. In press. Lazy Learning.
Jaap Murre & Peter Broeder (eds.), Cognitive Models of Language Acquisition. Cambridge, Mass.: MIT Press.

Gillis, Steven & Gert Durieux. In press. Learning Grammatical Classes from Phonological Cues. In Antonella Sorace, Caroline Heycock and Richard Shillcock eds. Language Acquisition: Knowledge Representation and Processing. Edinburgh: Edinburgh University Press.

Gleitman, Lila R., Harvey Gleitman, Barbara Landau & Eric Wanner. 1988. Where learning begins.
Frederick J. Newmeyer (ed.), The Cambridge Linguistic Survey, Vol. 3, 150-193. New York: Cambridge University Press.

Kelly, Michael H. . 1992. Using sound to solve syntactic problems. Psychological Review 99:349-364.

Kelly, Michael H. . 1996. The role of phonology in grammatical category assignment.
J. Morgan & K. Demuth (eds.), From Signal to Syntax, 249-262. Hillsdale: Erlbaum.

MacWhinney, B., J. Leinbach, R. Taraban & J. McDonald. 1989. Language learning: Cues or rules? Journal of Memory and Language 28:255-277.

Maratsos, Michael P. & Mary Anne Chalkley. 1980. The internal language of children's syntax.
Keith E. Nelson (ed.), Children's Language, Vol. 2, 127-214. New York: Gardner Press.

Mintz, T., E. Newport & T. Bever. 1995. Distributional regularities in speech to young children. 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, 263-283. Hillsdale: Erlbaum.

Pinker, Steven. 1987. The bootstrapping problem in language acquisition.
Brian MacWhinney (ed.), Mechanisms of Language Acquisition, 399-441. Hillsdale: Erlbaum.

Quinlan, J. Ross. 1986. Induction of decision trees. Machine Learning 1:81-106.

Weiss, Sholom M. & Casimir A. Kulikowski. 1991. Computer systems that learn. San Mateo: Morgan Kaufmann.

Wettschereck, Dietrich. 1995. A Study of Distance-Based Machine Learning Algoritms. PhD. Dissertation, Oregon State University.