In Proceedings of the Sixth International Conference on Cognitive Modeling, 358-359. Pittsburgh, Mahwah, NJ: Lawrence Earlbaum.

A Memory-Based Learning Model of Dutch Plural Inflection

Emmanuel Keuleers (Emmanuel.Keuleers@ua.ac.be)
Center for Psycholinguistics, University of Antwerp
Antwerp, Belgium

Dominiek Sandra (Dominiek.Sandra@ua.ac.be)
Center for Psycholinguistics, University of Antwerp
Antwerp, Belgium

A common observation in studies of inflectional morphology is that words which are atypical members of their syntactic class often take a regular inflection, independent of their phonology. For example, in English we would say that the plural of the surname MANN is MANNS although it has the same phonology as the noun MAN which has the irregular plural MEN. The same goes for the borrowed word TALISMAN (plural TALISMANS) as opposed to the noun FIREMAN (plural FIREMEN).

Marcus et al. (1995), refer to these atypical words as *non-canonical roots*, and suggest that their inflection can best be explained by a dual-mechanism model. Irregular words are inflected through an associative memory system, while regular words are inflected by a symbolic rule (e.g. add *-s).* For non-canonical roots however, access to the associative memory sytem is blocked because they are not marked as nouns or verbs. Consequently, they are inflected by the symbolic rule by default.

The inflection of non-canonical roots presents a problem for single-mechanism models such as the connectionist model for the English past tense proposed by Rumelhart and McClelland (1986). As words in these models are coded in terms of their phonological characteristics, atypical words cannot be distinguished from other words. Adding a feature that marks words as non-canonical entries could resolve this, but as Pinker and Ullman (2002) point out, this would not be very different from the implementation of a symbolic rule. To present a complete alternative to the dual-mechanism model, a single-mechanism associative model should be able to resolve the inflection of non-canonical roots based on characteristics that are not reducable to a symbolic rule.

Using simulations on corpus data, we will try to show that a single-mechanism model can inflect a particular type of non-canonical roots, namely unassimilated borrowings in Dutch, by using orthographic information. The choice for this type of information is grounded in the observation that the spelling of borrowings often deviates from typical Dutch spelling. If the plural of existing borrowings is stored along with their spelling pattern, this orthographic information could drive an analogical process with these stored representations and thus guide the selection of the plural.

Simulations

The dual-mechanism model takes a somewhat particular position with respect to the Dutch plural, in that both its productive plural suffixes (*-en* and *-s*) can be the default, dependent on the phonology of the root form (Pinker, 1999). This double default makes the Dutch plural a highly regular system: about 85 % of all words can be inflected by the default rule and little storage of exceptions is necessary. As non-canonical roots are inflected by the default rule, they should more or less reflect the distribution of the plural suffixes, taking -*en* more often than *-s.* This appears to be the case for some non-canonical roots (e.g. surnames), but as is commonly noted in descriptions of the Dutch plural system, unassimilated borrowings in Dutch tend to take the less frequent *-s* plural, which makes them an interesting test case. The goal for these simulations is to show that an associative model is able to correctly predict the plural of most words in a Dutch corpus, but that it is also able to predict the plural for unassimilated borrowings at least as well as the default part of a dual-mechanism model. This would be evidence that non-canonical roots can be stored and that their inflection is guided by analogical principles rather than by a default-rule.

To test our hypothesis, we compared the plural prediction accuracy of a phonological rule model (equivalent to the default part of the dual-mechanism model) to the accuracy of three single-mechanism models using diffferent information representations. The data-set on which the models were tested consisted of 3145 non-compound singular-plural noun pairs found in the CELEX lexical database for Dutch. The single-mechanism models were implemented using the Tilburg Memory Based Learner (TiMBL) (Daelemans et al., 2003), which can be seen as an extension to classical nearest neighbor models that allows for the specification of a number of parameters, such as number of neighbours, overlap and distance metric and feature weigthing. We made two changes to TiMBL's default parameter values: the modified value difference metric was used instead of the overlap metric and the number of neighbours participating in the class prediction was set to 10. All simulations used the leave-one-out method (for each word the plural was predicted once, using the rest of the data as a training set). Words were represented using the onset, nucleus and coda of the two final syllables. A first model (MBL-phon) used a purely phonological representation, while

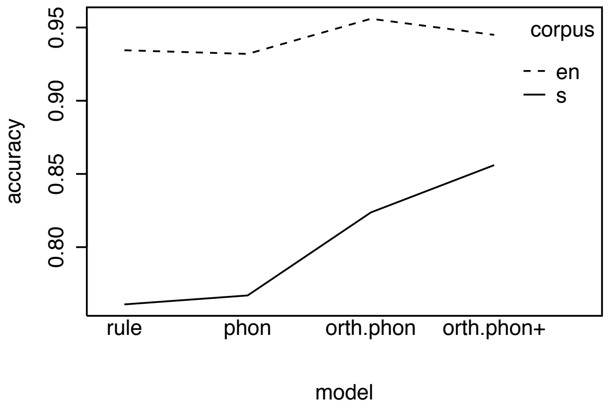

Figure 1: Accuracy of four models on predicting the plural of nouns from the CELEX lexical database.

a second model (MBL-orth-phon) used an additional orthographic representation. Finally, a third model (MBL-orth-phon+) was implemented using a representation where additional features for the distinctiveness of each of the orthographic features were derived using a sequential memory-based learning approach. The distinctiveness measures correspond to the number of phonological neighbors that do not share the orthographic feature. For example, in Dutch, the spelling *fr* is used in almost all words with the sound */fr/*, so its spelling is not distinctive. On the other hand, the spelling *ea* is almost never used in words with the sound */i:/*, so its spelling is highly distinctive. These additional features should give the MBL-orth-phon+ model the potential of finding neighbours that are not very similar in spelling or sound but that are similar in the distinctiveness of their spelling pattern.

Results and Discussion Figure 1 illustrates the results of the simulations. We performed a logit-analysis on prediction accuracy taking the rule model as the baseline. For words which had an *-en* plural in the corpus, neither the MBL-phon nor the MBL-orth-phon+ model were more successful than the Rule model. The MBL-orth-phon model on the other hand, was significantly better at predicting these plurals, $z = 2.970, p < .01$. For words which had an *-s* plural in the corpus, both the MBL-orth-phon ($z = 3.696, p < .001$) and MBL-orth-phon+ model ($z = 5.7335, p < .001$) performed significantly better than the Rule model.

These results show that an associative model can be successfully used to predict the suffix of Dutch plurals and that the addition of orthography as an information source constitutes a significant improvement . For *-s* plurals, where the Rule-model is successful in about 76 % of the cases, the MBL-orth-phon and MBL-orth-phon+ model offer a 6% and 10 % improvement respectively. The fact that the MBL-phon model does not do any better than the Rule model, indicates that this improvement is due to the additional information source, orthography.

Table 1: Words for which the *-s* plural was predicted correctly by the MBL-orth-phon and MBL-orth-phon+ models but for which the Rule model and MBL-phon predicted an *-en* plural.

BACK BARBECUE BIDON BOARD BOY CAKE CAPE CAPUCHON
CLAN CLOWN CLUB COAT COUPE CRACK CREPE CROSS DRIVE
E ECHELON ESKADRON FILE FRAME FRITE G GAME GOAL HALL
HOLE HOME INCH JACK JOULE KAMELEON KARTEL KICK L
MISS MOVE P PASS PIECE PLAQUE POLL R RACE RAID RELIEF
RIFF SHERIFF SHOP SHOW SNACK SPIKE SPRAY STICK STOCK
TAKE TAPE TIC TOUR TRACK TRICK TRUC TRUCK VUE WAGON
Y YANK

The question at this point is whether orthographic information is also used in the inflection of unassimilated borrowings, which under the dual-mechanism model are inflected by the default rule. As the Rule model is an implementation of the default it should always inflect these words correctly. Words that are inflected correctly by the MBL-orth-phon(+) model but incorrectly by the Rule model should therefore not include unassimilated borrowings or other non-canonical roots. Moreover, if it is to be shown unequivocally that orthography is the information source which allows for the correct inflection of these words, they should also be incorrectly predicted by the MBL-phon model.

As the appreciation of what constitutes an unassimilated borrowings is mostly qualitative, table 1 lists all the words for which the MBL models using orthography predicted an *-s* plural but for which both the Rule model and the MBL-phon model incorrectly predicted an *-en* plural. These constraints do not make for a very long list, but as can be observed, it contains almost exclusively English words, i.e. non-canonical roots in Dutch. The fact that these words are captured by an associative model using information that is not reducable to a symbolic rule, suggests that associative memory is not blocked for these words and that a default mechanism is not needed to inflect them. The failure of the default rule to inflect these words, poses an additional problem to the dual-mechanism model.

References

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2003). Timbl: Tilburg memory based learner, version 5.0, reference guide. Technical report, University of Tilburg.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, 29:189–256.

Pinker, S. (1999). *Words and Rules*. London: Phoenix.

Pinker, S. and Ullman, M. T. (2002). The past-tense debate: The past and future of past tense. *Trends in Cognitive Sciences*, 6(11):456–463.

Rumelhart, D. and McClelland, J. (1986). On learning the past tenses of English verbs. In *Parallel Distributed Processing, vol 2.* MIT Press.