The Learning of Weak Noun Declension in German: Children vs. Artificial Network Models

Peter Indefrey
Max-Planck-Institute for Psycholinguistics,
Postbus 310, NL-6500 AH Nijmegen, The Netherlands,
indefrey@mpi.nl

Rainer Goebel
Department of Psychology, University of Braunschweig,
Spielmannstr.19, D-3300 Braunschweig, FRG,
I3160905@DBSTU1.bitnet

Abstract

Different artificial networks are presented with the task of learning weak noun declension in German. This morphological rule is difficult for cue-based models because it requires the resolution of conflicting cue-predictions and a dynamic positional coding due to suffixation. In addition to that its 'task frequency' is very low in natural language. This property is preserved in the training input to study the models' abilities to handle low frequency rules. The performances of three kinds of networks:
1) feedforward networks
2) recurrent networks
3) recurrent networks with short term memory (STM) capacity
are compared to empirical findings of an elicitation experiment with 129 subjects of ages 5-9 and adult age.

Introduction

Connectionist network models have challenged the view that linguistic rules are also mentally represented in the form of symbolic rules. It is argued that representations based on subsymbolic cues show a good performance on language related tasks such as learning the English past tense (Rumelhart & McClelland 1987) or the declension of the German definite article (MacWhinney et al.1989), suggesting that cue-based representations might be good models for human language representation.

From the perspective of language acquisition these models are interesting because it is acknowledged by their authors that 'good performance' can not only mean a high percentage of correct responses but must also mean the correct modelling of children's behaviour in acquiring the linguistic rule in question. Phenomena resembling attested learner's errors such as overgeneralization of the past tense rule to irregular stems or omission of the definite article have indeed been found for the two aforementioned models. Whether these can be interpreted as modelling human language learner's behaviour, however, is doubtful for different reasons. Thus it is highly questionable whether the failure of the model of McWhinney et al.(1989) to reach critical activations of output nodes for a number of items is equivalent to the omission of articles in children's speech. The inability to select a particular article form is not the only possible cause of article omissions by children. They also have to learn that the definite article is syntactically obligatory in certain contexts - an intervening variable that was not represented in the model but is of great importance for the interpretation of the child language data. Additional variables on the model side can also be detrimental for comparability. Pinker & Prince (1988) showed that the U-shaped curve of overgeneralizations found for the model of Rumelhart & McClelland (1987) can not be taken as equivalent to the same curve found for human language learners for being an artifact of the timing and composition of the training input. They also pointed to the fact that a model's input representation has an effect on

the possible output, which is not only true for 'wickelfeature' coding but also for the positional coding used by MacWhinney et al.(1989) that is required by their feedforward network in order to generalize over fixed positions.

The present study compares the learning of a morphological rule of German - weak noun declension - by children to the learning of the same rule by different connectionist networks. It is designed to avoid effects of syntax acquisition and acquisition of the ability to detect the relevant cues as well as effects on the performance of the model caused by inappropriate or deliberately altered training input. The effect of different input representations is analysed.

We do not want to present ONE network solution but to evaluate the performance of different models with respect to some of the empirical acquisition data. The first goal of this comparison is to analyse whether cue-based models are in principle capable to simulate the acquisition process. The second goal is to study recurrent networks. These networks provide the possibility to sequentially present successive phoneme-codes. However, these models are limited to learn only short-ranging dependencies. The third goal is to study a more complex network architecture. As argued earlier (Goebel, 1990, 1993) we assume that the power of human information processing rests in a fruitful mixture of cue-based ('associative') processing and symbol manipulation abilities. We further assume that selective attention and short-term memory are necessary conditions for human symbolic manipulation. The short-term retention of information and the selective access of parts of that information - independent of its content - are mechanisms which are needed to successfully handle compositionality and provide a structure-sensitive processing capacity (for a detailed argumentation, see Goebel, 1990; a network model of symbolic manipulation using visually presented stimuli is described in Goebel, 1993).

Weak noun declension in German

Masculine nouns ending on '-e' (schwa) form the core of the weak declension class in German. In contrast to all others, nouns of this class receive a case ending '-(e)n' in dative and accusative singular. Although it is a rather small class - the core group consists of some 100 words referring to animals and human beings - it is nonetheless regular. Its properties make it an interesting test case for cue-based models. Both cues, gender 'masculine' and ending on '-e', in isolation strongly predict no case ending. This is because a) all other masculine nouns are strongly declined, which means that they receive no case ending in the singular except for genitive '-s'.
b) ending on '-e' is a powerful cue to feminine gender and feminines do not receive any case ending at all in the singular. Thus a cue-based model must learn to ignore the predictions of the single cues whenever they appear in combination. In addition to that the small size of the weak declension class (less than 1% of all nouns) brings about a very low task frequency, making it probable that such nouns will be permanently treated as exceptions by most networks (but possibly also by human learners).

Weak noun declension is acquired rather late, so that there is no interference with the acquisition of case as a syntactic category (by age 4; Clahsen, 1988). Neither is there any interference with the development of the ability to detect the relevant cues animacy, sex, ending on '-e', and gender (by age 3; MacWhinney, 1978, Schneuwly, 1978, Mills, 1984,1985)

The experiment

Procedure. A picture elicitation technique based on contrastive presentation of known and invented individuals or objects was developped. There were 12 fantasy pictures, each with a different combination of gender and one of the four semantic attributes 'inanimate', 'animate', 'animate + female', 'animate + male'. Fantasy pictures were presented simultaneously with pictures of known individuals of different genders (e.g. der Riese 'giant', die Biene 'bee'). In order to avoid any metalinguistic awareness the experimenter didn't model any examples. Subjects first repeated article and name of the individuals or objects presented, then a first accusative form was elicited faking that subjects were still in the repetition phase and asking them "Worauf zeige ich jetzt?" (What do I point to now?), suggesting an answer which makes use of the prepositional construction "auf den/die/das ...." (to the ...) with obligatory accusative case marking. Then plural was elicited by presenting two pictures with pairs of the same individuals. Finally two pictures were

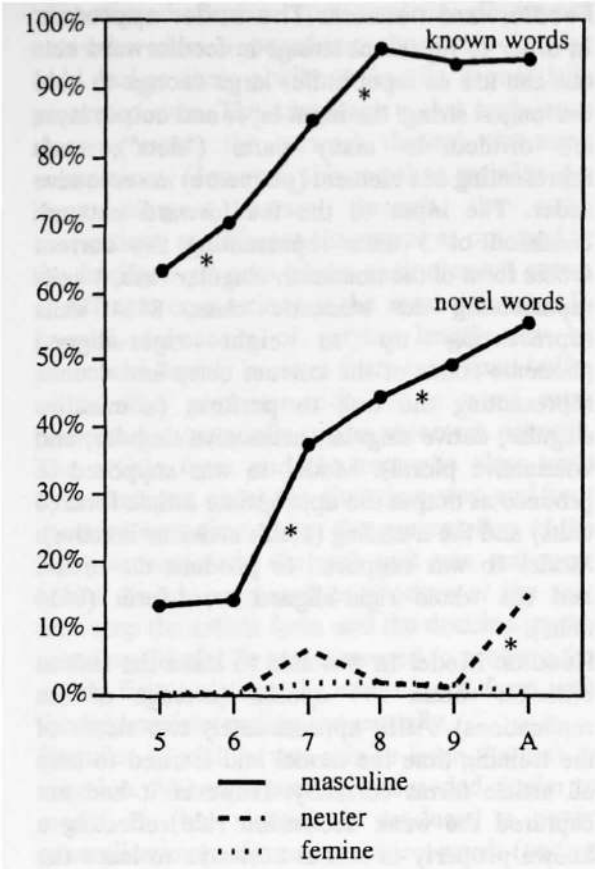

Figure 1. Relative number of n-endings for different ages (A = adults; * = significant, p=0.05)

presented in which the individuals were given to each other in reversed order by a third neutral individual ('the professor'). Recipient and given object receive dative resp. accusative case in German. Elicitation of plural, which could be considered fully acquired for the age groups in question, served as a baseline for productivity. Five groups of children aged 5 to 9 and two adult groups of different educational levels were tested. In total there were 129 subjects. The experimental sessions were recorded on DAT audio tape.

Results. A log-linear analysis was applied to the data. With respect to the network models the following results were most interesting:

1) Weak declension is productive and obligatory for masculine nouns ending on -e.

2) For known words it is fully acquired by age 8. (see figure 1, upper curve)

3) Learning of weak inflection starts out with rote learning.

4) There is a major shift towards productivity between age 6 and 7. (see figure 1, middle curve)

5) There are no overgeneralizations to feminine nouns or - with the exception of adults - neuter nouns. (see figure 1, lower curves)

6) The effect of picture semantics on weak declension is marginal and unsystematic.

The models

The training corpus

The complete active vocabulary of 10 children (5;6 - 6;0) collected over a period of four months (Augst, 1984) was taken as a basis for the construction of the training corpus. Of the total 15000 different word types there were 6364 noun types, of which less than 1% of the weak declension class. For every noun its frequency was estimated on the basis of its lexical dispersion, that is how many of the 10 children used it, applying the formula:

$$(1) \qquad f = -k \ln (1 - d/k)$$

where f = frequency,d = dispersion,k = number of children (Indefrey & Baayen, 1993). From the frequency-weighted corpus a training corpus of 1000 noun types (1349 tokens) was drawn at random. It contained 13 (=1,3%) masculine nouns ending on '-e' with a total of 21 (=1,6%) tokens. Case and number inflected forms of the corpus entries were phonetically transcribed and feature coded following Wurzel (1981). Each phoneme was represented with 14 units. In addition every entry contained a frequency weight, case and number inflected forms of the definite article and a code for one of the four semantic classes 'inanimate', 'animate', 'animate + male', 'animate + female'. Nominative, dative and accusative singular as well as nominative plural definite forms (e.g. dem Jungen) were trained according to their relative frequency in spoken language.

The networks

The performance of three kinds of networks on

two different tasks was analysed:

1) feedforward network
   (with and without hidden units)
   input:
   phoneme-codes presented simultaneously;
   right-aligned word ending
   output:
   a) article plus yes/no decision on n-ending
   b) whole word-form, right-aligned
2) recurrent network
   input:
   phoneme-codes presented sequentially
   output:
   a) article plus yes/no decision on n-ending
   b) sequentially produced phoneme-codes of whole word
3) STM-based recurrent network
   input:
   phoneme-codes presented sequentially and stored in STM
   output:
   sequentially produced phoneme-codes;
   phoneme-codes may be accessed sequentially from STM

Training procedure. Each network was trained on the whole corpus and interrupted at several steps during learning based on the overall error value. At each stage the performance of the network was tested on the training corpus. Then generalization performance was tested on the set of novel words that was used for the child experiment. Each individual simulation was replicated ten times using each time a new set of random weights drawn from the interval -1, 1.

Technical details. The networks were trained using the back-propagation learning rule (Rumelhart, Hinton & Williams, 1986). In order to apply the generalized delta rule to the recurrent networks the "unfolding through time" technique was adopted (Minsky & Papert, 1969; Rumelhart, Hinton & Williams, 1986) and improved (Williams & Zipser, 1990). In order to speed up learning a modified version of Fahlman's (1988) "quick-prop algorithm" was used. In all simulations, the learning criterion was set such that the activity value of each output unit should not deviate more than 0.1 from the supplied target value. The target values for the studied recurrent network could be specified as 'don't care values' (Jordan, 1986) meaning that an output unit is allowed to produce an arbitrary value at a certain time step.

Feedforward network: The buffer approach. In order to represent strings in feedforward nets one can use an input buffer large enough to hold the longest string: the input layer and output layer are divided in many parts ("slots"), each representing one element (phoneme) in successive order. The input to the feedforward network consisted of 3 units representing the current article form of the nominativ singular case, 4 units representing the semantic class, 8*14 units representing up to eight right-aligned phoneme-codes of the current entry and 4 units representing the task to perform (nominative singular, dative singular, accusative singular, and nominative plural). Model 1a was supposed to produce as output the appropriate article form (6 units) and the n-ending (1 unit active or inactive). Model 1b was supposed to produce the article and the whole right-aligned word-form (8*14 units).

Results. Model 1a was able to learn the task to criterion within 174 epochs (average of ten replications). After approximately two thirds of the training time the model had learned to map all article forms correctly. However it had not captured the weak declension rule reflecting a known property of neural networks to learn the strongest regularities first. Within the final third of training time the model gradually acquired the weak declension rule. An analysis of the weight pattern showed that the model indeed used the appropriate cues (ending on '-e' and masculine gender) but it also used the specific phoneme-codes of the relevant words which can be interpreted as rote learning. This had an impact on generalization performance: although all test words ending on '-e' and masculine gender stimulated the 'n-ending' output unit correctly the amount of activation did not reach a value greater than 0.5 in half of the test cases.

Model 1b was unable to learn the task to criterion. Although it learned the correct article forms the model had great difficulties in learning the (right-aligned) output word form. The problem is that if a mapping requires an ending (e.g., '-n' or '-s') at the output layer the phomeme-codes at the input layer have to be mapped shifted one position to the left at the output layer.

Recurrent Network: Sequential Processing. A disadvantage of feedforward networks using the buffer approach is that acquired knowledge is position dependent, thus it was necessary to use a

right-aligned input representation. Recurrent networks seem a promising alternative for the task due to their ability to handle sequentially presented data. The advantage is that each input element enters the network through the same connections. However, this requires seriality: An input element enters through the same connections at different time steps as opposed to the feedforward case where each element enters at different connections at the same time step. In general, sequences of varying length may be processed flexibly because there is no fixed buffer width. The phoneme-codes of a word were presented sequentially to the recurrent network. The article form and the semantic class units were clamped on at the first time step and kept constant over time. After the presentation of the last phoneme-code the task unit was activated. Model 2a was supposed to produce at the next time step the article form and the decision on the n-ending. Model 2b was supposed to produce the article first and then the whole word form with the appropriate ending sequentially.

Results. Model 2a was able to learn the task to criterion. Although learning proceeded similar to model 1a (but slower), it produced a better generalization performance: all test words (ending on '-e' and masculine gender) produced an activation level of the n-ending unit which exceeded 0.5. The reason for this seems to be that the model can not exploit long-ranging dependencies for rote learning as easily as the feedforward network. Model 2b did not learn the task to criterion. The task is much more difficult to solve than the one for model 2a since the recurrent network must learn to construct an appropriate compact internal representation of the sequentially presented elements which enables the network to convert this representation back to a correct output sequence. The network has to construct an appropriate internal 'plan vector' (see Jordan, 1986). This is a difficult problem since the delay between the production of the word ending and the critical information consists of several time steps and the injected error values are propagated back through a deeply layered (unfolded) network.

STM-recurrent model. (see figure 2) The sequentially presented input word is routed to two subsystems, one short-term store and one recurrent network similar to model 2b. The short-term store consists of a recurrent network with fast weights allowing to hold a novel sequence for some time (for details of the construction of the short-term store, see Goebel, 1990). The model operates in two stages, a recognition stage and a production stage. In the first stage, a sequence is presented and the network is trained to recognize it at a locally represented word-form layer (corresponding roughly to a lexical entry). If it is a new word-form the most similar entry may be activated. In the production stage both the activated word (acting as a 'plan') together with the sequentially accessible STM representation are used to produce a sequence of phoneme-codes at the output layer. The network has the possibility to exploit both the sequentially 'rehearsed' phoneme-codes (as model 2) and the activated lexical entry node to produce the output phoneme-sequence. The availability of the STM-representation is especially important during learning since its access is very helpful for producing the right output elements. If during testing a novel word is presented, the most similar entries may be partially activated. The generalization performance depends then on both, the representation at the lexical entry layer and the representation held in STM. The model is in agreement with the framework proposed by Pinker (1991) since there are two subsystems, one for 'regular rules' (exploiting STM), the other

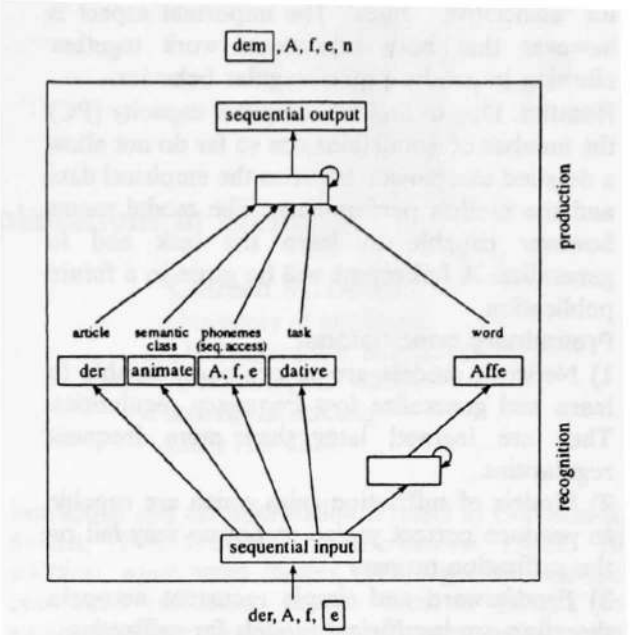

Figure 2. Structure of STM recurrent network. (Phonemes for typographic reasons not in IPA notation)

for 'associative rules'. The important aspect is however that both subsystems work together allowing to produce quasi-regular behavior.

Results. Due to limited computer capacity (PC) the number of simulations run so far do not allow a detailed comparison between the empirical data and the models performance. The model seems however capable to learn the task and to generalize. A full report will be given in a future publication.

Preliminary conclusions.

1) Network models are in principle capable to learn and generalize low frequency regularities. They are learned later than more frequent regularities.

2) Models of suffixation rules which are capable to produce correct yes/no responses may fail on the suffixation proper.

3) Feedforward and simple recurrent networks therefore are insufficient models for suffixation.

4) For all models there was more evidence for a gradual improvement of performance on the task than for a major learning step like the one found empirically.

References

Augst, G. 1984. Kinderwort. Frankfurt a.M.: Verlag Peter Lang.

Clahsen, H. 1988. Normale und gestörte Kindersprache. Amsterdam: John Benjamins

Fahlman, S.E. 1988. Faster Learning Variations on Back-Propagation: An Empirical Study. In: Touretzky, D., Hinton, G. and Sejnowski, T. eds. Proceedings of the 1988 Connectionist Models Summer School.

Goebel, R. 1990. Binding, Episodic Short-Term Memory and Selective Attention, Or Why are PDP Models Poor at Symbol Manipulation? In: Touretzky, D.S., Elman, J.L., Sejnowski T.J. & Hinton G.E. eds. Connectionist Models. Proceedings of the 1990 Summer School. San Mateo: Morgan Kaufman.

Goebel, R. 1993. The role of visual perception, selective attention, and short-term memory for symbol manipulation: A neural network model that learns to evaluate simple LISP expressions. In: Wender K.F., Schmalhofer F. & Boecker H.D. eds. Cognition and Computer Programming. Ablex Publishing Corporation, in press.

Indefrey, P. and Baayen, H. 1993. Estimating Frequencies From Lexical Dispersion Data. Forthcoming.

Jordan, M.I. 1986. Attractor dynamics and parallelism in a connectionist sequential machine. Proceedings of the Eighth Annual Conference of the Cognitive Science Society, 531-546.

MacWhinney, B. 1978. The Acquisition of Morphophonology. Monographs of the Society for Research in Child Development, Vol.43

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

Mills, A.E. 1984. The Acquisition of Gender in English and German. Habilitationsschrift, Tübingen

Mills, A.E. 1985. The Acquisition of German. In: Slobin, D. ed. The Cross-linguistic Study of Language Acquisition. Hillsdale, New Jersey

Minsky, M. and Papert, S. 1969. Perceptrons. Cambridge, MA: MIT Press.

Pinker, S. and Prince, A. 1988. On language and connectionism: Analysis of a parallel distributed model of language acquisition. Cognition, 28, 73-193

Pinker, S. 1991. Rules of Language. Science, 253, 530-535.

Rumelhart, D.E., Hinton, G. and Williams, R.J. 1986. Learning Internal Representations by Error Propagation. In Rumelhart, D.E. and McClelland, J.L. eds. Parallel Distributed Processing. Volume I, MIT Press, Cambridge

Rumelhart, D.E. and McClelland, J.L. 1987. Learning the Past Tenses of English Verbs: Implicit Rules Or Parallel Distributed Processing. In: MacWhinney, B. ed. Mechanisms of Language Acquisition. Hillsdale, New Jersey

Schneuwly, B. 1978. Zum Erwerb des Genus im Deutschen: eine mögliche Strategie. Unpublished Manuscript, Max-Planck-Institut für Psycholinguistik, Nijmegen

Williams, R.J. and Zipser, D. 1990. Gradient-Based Learning Algorithms For Recurrent Connectionist Networks. Technical Report NU-CCS-90-9. College of Computer Science.

Wurzel, W.U. 1981. Phonologie: Segmentale Struktur. In: Heidolph, K.E. et al. eds. Grundzüge einer deutschen Grammatik. Berlin: Akademie-Verlag, 898-990