In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society, 253-258*, Lawrence Erlbaum, Hillsdale, New Jersey, 1992

Integrating Category Acquisition with Inflectional Marking: A Model of the German Nominal System

Prahlad Gupta
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
prahlad@cs.cmu.edu

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
brian@andrew.cmu.edu

Abstract

Linguistic *categories* play a key role in virtually every theory that has a bearing on human language. This paper presents a connectionist model of grammatical category formation and use, within the domain of the German nominal system. The model demonstrates (1) how categorical information can be created through co-occurrence learning; (2) how grammatical categorization and inflectional marking can be integrated in a single system; (3) how the use of co-occurrence information, semantic information and surface feature information can be usefully combined in a learning system; and (4) how a computational model can scale up toward simulating the full range of phenomena involved in an actual system of inflectional morphology. This is, to our knowledge, the first connectionist model to simultaneously address all these issues for a domain of language acquisition.

Introduction

In virtually every model of language processing, the notion of linguistic category plays a key role. For example, syntactic categories such as *noun* and *verb* are the stuff of which sentence processing is thought to be made; grammatical categories such as *gender* and *person* are essential to the co-ordination of conjugational and declensional paradigms in many languages. Linguistic categorization has thus usually been a cornerstone of thinking about language.

This paper presents a connectionist account of how grammatical categories could be formed and usefully incorporated into processing. The phenomenon we model is learning of grammatical gender, within the German nominal system. This domain involves coordination of case, number, and gender information, and for this reason has often been regarded as a challenge to models of language acquisition (Maratsos and Chalkley, 1980; Maratsos, 1982; Pinker, 1984). We therefore chose this domain as an excellent test-bed for proposals about cue-driven learning and categorization.

Aims and Relation to Previous Work

A number of models of linguistic category acquisition have previously been proposed (MacWhinney, 1978; Maratsos and Chalkley, 1980; Pinker, 1984). The similar accounts in (MacWhinney, 1978) and (Pinker, 1984) both involve row- and column-splitting algorithms that operate on a data structure representing the paradigm for the German definite article. However, these matrix-manipulation operations are rather *ad hoc* in nature; problems with these accounts are discussed in more detail in (MacWhinney, 1991). The account in (Maratsos and Chalkley, 1980) and (Maratsos, 1982), while intuitively appealing, has not been specified in computationally precise form.

The aim of the present research was to provide a computational account of the formation of the grammatical category of gender in German, and of how this categorical information could be usefully employed in language processing and acquisition, without reliance on the kinds of *ad hoc* mechanisms specified in the earlier MacWhinney-Pinker account. We aimed, moreover, to make this computational investigation within a connectionist framework.

Previous work by the second author and colleagues has presented a computational model of the acquisition of the German definite article (MacWhinney, Leinbach, Taraban and McDonald, 1989; Taraban, McDonald and MacWhinney, 1989). As will be discussed in more detail in the final section, the present work achieves several significant advances over the earlier model, while also replicating the earlier results.

The German Nominal System

The system of *grammatical gender* in German assigns every noun to one of three gender categories: *masculine, feminine*, or *neuter*. The grammatical gender assigned to a noun will in general have little to do with the sex of its referent. For example, the noun *Fräulein*, meaning "young lady", has neuter gender, while the noun *Polizei*, meaning "police", has feminine gender.

As shown in Table 1, the correct definite article for use with a given noun depends on the gender of the noun, and on the case and number in which the noun is

Case	Singular			Plur
	Masc	Fem	Neut	
Nominative	der	die	das	die
Genitive	des	der	des	der
Dative	dem	der	dem	den
Accusative	den	die	das	die

Table 1: Gender, number and case paradigm for the German definite article.

used. Potentially, this leads to 24 cells in the paradigm (4 case possibilities x 3 gender possibilities x 2 number possibilities). However, gender is not relevant in the plural number, and so there are only 16 cells in the paradigm. As there are only six distinct definite articles (*der, die, das, des, dem, den*), a particular article obviously can and does appear in more than one cell.

The stem of a noun undergoes various inflectional modifications according to the case and number context in which it is used, and also depending on its gender. Possible inflectional changes include *umlauting* of a vowel in the stem, and various *suffixation* processes, with *voicing* of a final consonant accompanying certain suffixes. These changes are discussed in more detail in (Mugdan, 1977).

The Model

The architecture of the model is shown in Figure 1. Essentially, this is a connectionist architecture, though with some departures from what is most typical of such models. The overall system consists of three networks, described below: a *categorization network*, an *article-learning network*, and a *stem-modification-learning network*.

Categorization Network

The *categorization network* is shown in the region marked 1 in Figure 1. It constitutes a mechanism that learns to *categorize* articles, based on their co-occurrence with case and number information. This takes the form of a *competitive learning* network (Rumelhart and Zipser, 1986) whose inputs are the representations of case, number, and the article, and whose output response is a pattern over the "Winner-Take-All" layer that identifies that case-number-article combination.

We have assumed that there is a "lexicon", consisting of "lexical representations" of noun stems[1]. For our current purposes, a "lexical entry" comprises information about both the phonology of the noun and the co-occurrence relations in which the noun has participated. In the present case, this latter information is limited to co-occurrences with particular articles. We assume that the categorization responses of the competitive learning network shape the part of the lexical representation of the noun that stores co-occurrence information. Over time, this lexical information comes to be a trace of which articles have occurred with the noun in which case and number. These encodings constitute the noun's *co-occurrence history*.

There are fourteen possible distinct combinations of Case, Number and Article that can occur. These are: Nom-Sing-der (Nominative-Singular-der), Gen-Sing-des, Dat-Sing-dem, Acc-Sing-den, Nom-Sing-die, Gen-Sing-der, Dat-Sing-der, Acc-Sing-die, Nom-Sing-das, Acc-Sing-das, Nom-Plur-die, Gen-Plur-der, Dat-Plur-den, and Acc-Plur-die.

Competitive learning results in single, specific units in the Winner-Take-All layer responding to each possible combination. Note that the Winner-Take-All layer consists, not of exactly fourteen predetermined units, but of an arbitrary number of units (we used 50). Nevertheless, the unsupervised competitive learning algorithm results in there being fourteen units that come to "recognize" the fourteen possible combinations[2].

Only certain combinations of case, number and article will co-occur with a noun of a particular gender. For example, for a Feminine noun such as *Frau*, only the combinations Nom-Sing-die, Gen-Sing-der, Dat-Sing-der, Acc-Sing-die, Nom-Plur-die, Gen-Plur-der, Dat-Plur-den, and Acc-Plur-die will be observed; Feminines will not co-occur with Nom-Sing-der or Gen-Sing-der. Thus, a certain set of combinations of case, number and article will co-occur with Feminine nouns, a different set with Masculine nouns, and a different set for Neuter nouns.

It is important to note that articles are homophonous. For example, *der* is used with both Masculine and Feminine nouns. Occurrence of a particular article with a particular noun therefore does not provide sufficient information to determine the noun's gender (except for the article *das*). The *set* of all articles that can occur with a particular noun does provide sufficient information to encode gender uniquely. So also does the set of all possible combinations of case, number and article. However, if only *part* of the paradigm for a noun has been observed, then a record of the observed case-number-article combinations is a more robust encoding of gender than a record of only the observed articles.

In the model, the co-occurrence history for a particular noun stem is formed in the following way. The cate-

[1] Although, for convenience, we have depicted the lexicon as an array-like data structure, we envisage it as a collection of topographically organized maps. We have not attempted to implement this lexical organization; however, work by Mikkulainnen has demonstrated how such a *distributed lexicon* could be formed (Miikkulainen, 1990).

[2] The classification is sometimes into thirteen rather than fourteen categories, with the combinations Nom-Sing-das and Acc-Sing-das being grouped into a single category. However, this does not affect the usefulness of the categorizations to be discussed in the section on "Simulations and Results".

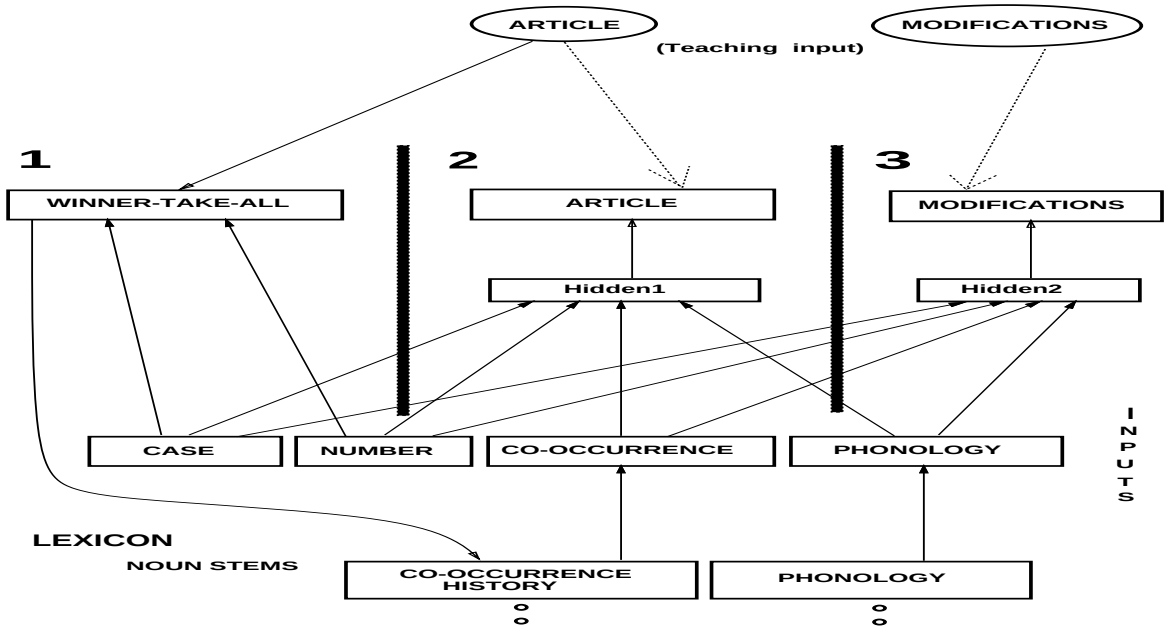

Figure 1: Architecture of the model used in simulations.

gorization responses for each case-number-article combination observed with the stem are additively encoded in the "co-occurrence history". This additive encoding involves the arithmetic addition of the pattern of activation evoked over the Winner-Take-All layer to the co-occurrence history part of the noun stem's lexical representation. As successive categorization responses are added to a particular noun's co-occurrence history, additional units in the co-occurrence history come to be active. Recall that the sets of case-number-article combinations that can co-occur with nouns of different gender are different. Therefore, different sets of units will come to be active in the co-occurrence history of stems of different gender. In other words, the lexical co-occurrence history comes to form a distributed representation of the *grammatical gender* of the stem.

Article-learning and Modification-learning Networks

The *article-learning* and *modification-learning* networks are shown in the regions marked 2 and 3 in Figure 1. These networks together model the process by which the child could learn to use the cues of Case, Number, the phonology of the noun, and its co-occurrence history, to predict the correct article, as well as to produce the corrected inflected form of the noun stem. In what follows, we will sometimes refer to the combination of the article-learning network and the modification-learning network as the *inflectional system*.

Each of these two networks is a typical three-layer connectionist architecture, whose inputs are representations of the noun's case, number, phonology and co-occurrence history. Case is represented by an 8-bit vector in which each of the four case possibilities is coded for by two bits. Number is represented by a 4-bit vector in which each of the two number possibilities is encoded in two bits. The phonological input is a 216-bit vector consisting of phonological distinctive feature representations of each phone in the noun stem; for further details of the phonological representation, the reader is referred to (MacWhinney et al., 1989). The hidden layer of each of these two networks comprises 60 units.

The output of the article-learning network is a representation of the correct article. This representation is a 12-bit vector in which two bits encode each of the six possible articles.

The outputs of the modification-learning network are the appropriate modifications that must be made to a noun stem, for a particular case, number and gender. The nine possible stem modifications are: umlauting of a vowel; addition of one or more of the suffixes *-e*, *-n*, *-s*, *-r*, *-ina*, *-se*, and *-ien*; and voicing of the final consonant in certain cases of suffixation. The output is represented as a 9-bit vector with one bit encoding each of the nine modifications.

Note that more than one of these modifications may be applicable to a particular noun stem in a particular case and number[3]. The primary determinants of the

[3]Note also that, although the total number of possible modifications is small, selection of the appropriate *set* of modifications for a given stem in each of the the eight cells

correct set of nominal markings given a particular case and number include (i) gender, (ii) the details of the phonological form of the stem, and (iii) a variety of semantic features which are not included in the present model. A complete linguistic analysis of this system can be found in (Mugdan, 1977).

As an example of training, suppose that the phrase *die Männer*, meaning "the men" (nominative plural), has been "heard". The inputs to both the article-learning network and the modification-learning network are patterns of activation representing Nominative case, Plural number, the phonology of the noun stem *Mann*, and the co-occurrence history of articles with the stem *Mann*. The article-learning network is trained to associate these items of information with the article it has observed (*die*). At the same time, the modification-learning network is trained to associate these same inputs with the inflectional changes that must be made to the stem *Mann*, viz., umlauting of the vowel, and suffixation of *-er*.

Simulations and Results

In the absence of detailed information about the linguistic input available to children learning German, we have based our data sets on a corpus of over 80,000 words from adult German usage (Wangler, 1963). From this corpus, we selected (on the basis of frequency) 2,094 inflected forms of 1,234 noun stems as the training data set, and another 315 inflected forms as a test data set.

Each *trial* involved presentation of input representing one of the 2,094 training patterns to the categorization, article-learning and modification-learning networks[4]. One *epoch* consisted of a trial for each of the 2094 words in the training set.

During training, the article-learning network was trained to produce the article appropriate for the presented word, while the modification-learning network was trained to produce the stem modifications appropriate for the presented word. In both cases, training was via the back-propagation learning algorithm (Rumelhart et al., 1986a). Synchronously, on each trial, the categorization network was trained to categorize the co-occurrence of Case, Number and Article, via the competitive learning algorithm (Rumelhart and Zipser, 1986). This categorization response was additively encoded in the lexical representation of the noun stem, as described in the section discussing the categorization network. As a result, on next access of the lexical representation of this stem, the modified co-occurrence history became available.

	% errors in:			
Epoch	NOM	GEN	DAT	ACC
5	1 %	31 %	3 %	10 %
10	0 %	22 %	1 %	1 %
15	0 %	10 %	0 %	0 %
20	0 %	6 %	0 %	0 %

Table 2: Percentage of errors made by the article-learning network in various case contexts over the first 20 epochs of training. NOM=Nominative, GEN=Genitive, DAT=Dative, ACC=Accusative.

Simulation 1 was run exactly as described above. The article-learning network learned to produce the correct article for all 2094 patterns in the training set in 66 epochs of training. The modification-learning network learned to produce the correct stem inflections in 68 epochs of training.

The types of errors made by the article-learning network at early stages in learning (over the first 20 epochs) parallel those made by German children learning this paradigm. First, the network learned all nominative forms within 5 epochs of training (see Table 2), which corresponds to childrens' early acquisition of the nominative. Second, the network made errors on an average 17% of genitive forms per epoch over the first 20 epochs, which corresponds to childrens' delayed acquisition of the genitive. Both of these results can be explained in terms of the fact that our training set incorporated approximately the real-world percentages of occurrence of various cases (40% for nominatives, 10% for genitives). Third, the response produced by the network was often below threshold for any of the possible articles, which corresponds to childrens' omission of articles. Fourth, the most common error was production of *der* for *des* for masculine and neuter nouns in the genitive singular, which would have been correct had the noun been of feminine gender (see Table 1). This can be interpreted as paralleling the child's overgeneralizations of a particular gender. These aspects of childrens' errors on the definite article are discussed in (MacWhinney, 1978) and (Mills, 1986).

To test generalization abilities, we examined the responses of the networks to patterns on which they had not been trained. The testing set of 315 forms consisted of 175 forms representing stems the networks had been trained on in other case-number contexts (*familiar-stem tests*), and 140 forms representing stems the network had not been exposed to at all (*novel-stem tests*). Once the article-learning network had learned the training set with 100% accuracy, it produced an incorrect article on only 7 of the 175 familiar-stem test forms (4% error rate), and on only 14 of the 140 novel-

of the declension (i.e., in each of the eight possible case-number combinations) involves a complex set of conditions. German has a large number of declensional classes with different assignments across these eight cells, with each class composed of many subgroups, partial regularities, and lists of exceptions.

[4]As noted previously, these inputs were representations of the Case, Number, stem phonology, and stem co-occurrence history. During training, the correct article and stem modifications were also presented, whereas during testing, they were not presented.

stem test forms (10% error rate)[5]. Similarly, once the modification-learning network had learned the training set to criterion, it produced correct modifications on 257 of the 315 generalization test forms (82% correct generalization). Thus, both the article-learning and modification-learning networks exhibited a substantial capacity for both kinds of generalization.

Co-occurrence information was created as described in the section discussing the categorization network; it categorized stems according to gender. To examine the *usefulness* of this information, we ran a simulation (Simulation 2) in which the co-occurrence information was not provided to the article-learning and modification-learning networks. This simulation was in every other respect identical to the one previously described.

In Simulation 2, it took 800 epochs for the article-learning network to learn to produce the correct article for all items in the training set. This is significantly worse performance than that in Simulation 1 (error-free production of the article in 66 epochs of training). Furthermore, the errors made by the article-learning network at early points in training during Simulation 2 were mostly on nominative forms. This is quite unlike the developmental course observed in children, and also unlike Simulation 1. In Simulation 2, it took 117 epochs for the modification-learning network to learn to produce stem modifications correctly for the entire training set. This compares with 68 epochs in Simulation 1.

These comparisons between Simulations 1 and 2 demonstrate that the categorical grammatical gender information that develops is genuinely useful for processing, and highlights the fact that the explicit re-representation of information may be an important technique for models developed within the overall connectionist framework.

Discussion

The use of a separate competitive learning network appears to capture important categorization effects in the process of learning the German article. The question arises, however, of whether such processing has applicability outside the present domain. In this connection, it is interesting to note that the hippocampus has been hypothesized to create orthogonalized episodic encodings (McClelland et al., 1992), which is very similar to the notion of encoding co-occurrences in the present categorization network. The same general categorization mechanisms potentially also provide a basis for the encoding of various regularities and sub-groupings. For example, for the German nominal inflection system, such a mechanism could lead to lexical encodings of the *pluralization paradigm class* of the noun stem. Similarly, for a language such as Hungarian, co-occurrences could lead to encoding of the *vowel harmony class* of the stem. Thus mechanisms very similar to what we propose may, in fact, play an important and quite general role in learning and memory processes.

We have hard-wired the categorization network to receive only exactly the inputs that were expected to be powerfully predictive of gender, namely, Case, Number and Phonology. At present we do not have a satisfactory answer to this criticism, except to note that this criticism is probably partially applicable to almost any model that makes assumptions about input and output information. Further work would be needed to determine the performance of the categorization network under conditions of noisy and extraneous data.

As mentioned at the beginning of this paper, previous work by the second author (MacWhinney et al., 1989) has addressed some of the same issues as the present model. This earlier work presented a computational model that learned the definite article in German without rules, and which matched the developmental data. The present model also uses a cue-driven system to match the developmental sequence of article learning observed in German children.

However, in achieving our aim of modeling the formation and utilization of grammatical gender in German, we feel we have made the following additional, significant, demonstrations, none of which was addressed by the (MacWhinney et al., 1989) model.

First, we have demonstrated how categorical information can be *created* through co-occurrence learning, *made available* in explicit distributed form, and usefully *utilized* by other parts of the processing system. The categorizations created by the competitive learning network in our model in effect construct the paradigm for the German definite article, but without reliance on the problematic row- and column-splitting mechanisms in the MacWhinney-Pinker account (MacWhinney, 1978; Pinker, 1984). We have shown how the encoding of these categorizations in the lexicon can lead to classification of nouns by gender. We have also shown that this gender/co-occurrence information is useful in the processes of learning the inflectional system.

Second, we have combined the task of learning the German definite article, examined previously in (MacWhinney et al., 1989), with the task of learning to inflect the noun stem that the article accompanies. This is a significant extension over the earlier model, both in terms of *coverage of linguistic phenomena*, and in terms of *integration of different kinds of processing* (grammatical categorization, and inflectional marking). We are not aware of any previous computational model of the present domain that combines these processes.

Third, we have demonstrated the integration of vari-

[5] Non-erroneous responses consisted of either production of the correct article (78% and 58% for familiar- and novel-stem generalization, respectively), or omission of the article altogether (18% and 32% for familiar- and novel-stem generalization, respectively).

ous *types* of information that have been regarded as important in language learning, viz., co-occurrence information (co-occurrence of Case, Number and Article), semantic information (the semantically based notions of Case and Number), and surface features (phonological information), and we have shown how these types of information can be usefully combined in a learning system. In effect, we have devised a computational implementation of the type of learning proposed in (Maratsos and Chalkley, 1980). To the best of our knowledge, such an implementation has not previously been constructed.

Fourth, we believe that it is vital for cognitive modeling of language to scale up to dealing with realistically sized data sets, because it is only then that linguistic regularities and sub-regularities really emerge. Our simulations used over 1200 noun stems in over 2000 inflected forms. We feel that this steps beyond the realm of a toy-sized model, and thus constitutes the beginnings of an important demonstration of realistic robustness. It also represents a substantial scaling up from the model in (MacWhinney et al., 1989), which used a training corpus consisting of 305 inflected forms of 102 noun stems[6].

In conclusion, the present work offers the first computational account of the synthesis of various kinds of information that have been regarded as important in language leaning. It also suggests how grammatical categories could develop and constitute useful processing information. Finally, this research begins to address questions about the ability of models of language acquisition to scale up to dealing with more realistic data.

Acknowledgements

We thank Jay McClelland, Dave Plaut and Dave Touretzky for helpful discussion, and Jay McClelland for use of computing facilities. Jared Leinbach developed formatting programs for network inputs and outputs. Of course, the present authors remain responsible for any errors in this work.

References

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, 43.

MacWhinney, B. (1991). Connectionism as a framework for language acquisition theory. In Miller, J., editor, *Research on Child Language Disorders*. Pro-Ed, Austin, TX.

MacWhinney, B., Leinbach, J., Taraban, R., and McDonald, J. L. (1989). Language learning: Cues or rules? *Journal of Memory and Language*, 28:255–277.

Maratsos, M. (1982). The child's construction of grammatical categories. In Wanner, E. and Gleitman, L., editors, *Language Acquisition: The State of the Art*. Cambridge University Press, New York.

Maratsos, M. and Chalkley, M. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In Nelson, K., editor, *Children's Language*, volume 2. Gardner, New York.

McClelland, J., McNaughton, B. L., O'Reilly, R. C., and Nadel, L. (1992). Complementary roles of hippocampus and neocortex in learning and memory. Society for Neuroscience abstracts, submitted.

Miikkulainen, R. (1990). A distributed feature map model of the lexicon. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum, Hillsdale, NJ.

Mills, A. E. (1986). *The Acquisition of Gender*. Springer-Verlag, Berlin.

Mugdan, M. (1977). *Flexionsmorphologie und Psycholinguistik*. Gunter Narr, Tubingen.

Pinker, S. (1984). *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA.

Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning internal representations by error propagation. In (Rumelhart et al., 1986b).

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986b). *Parallel Distributed Processing*, volume 1: Foundations. MIT Press, Cambridge, MA.

Rumelhart, D. E. and Zipser, D. (1986). Feature discovery by competitive learning. In (Rumelhart et al., 1986b).

Taraban, R., McDonald, J. L., and MacWhinney, B. (1989). Category learning in a connectionist model: Learning to decline the German definite article. In Corrigan, R., Eckman, F., and Noonan, M., editors, *Linguistic Categorization*. Benjamins, New York.

Wangler, H. H. (1963). *Rangworterbuch hochdeutscher Umgangsprache*. Elwert, Marburg, Germany.

[6] We have limited our data set to approximately 2,000 training forms, in order to reduce the time required to run a simulation. (Larger training sets would mean more stimuli per epoch, but would not affect the computational *tractability* of the simulation). However, it is not clear at what training set size all the basic regularities and patterns will be represented in the input set. We therefore consider it important to examine the effect of further increases in training set size.