

A Symbolic Model for Learning the Past-Tenses of English Verbs

Charles X.F. Ling Steven Cherwenka

Department of Computer Science
University of Western Ontario
London, Ontario, Canada N6A 5B7
Email: ling@csd.uwo.ca

Marin Marinov

Department of Philosophy
Sofia University
15 Ruski Bd., Sofia-1000, Bulgaria

Abstract

Rumelhart and McClelland first studied and implemented an artificial neural network (ANN) model for learning the past tenses of English verbs, and it became a landmark test for ANNs in language learning. Their model received widespread criticism from Pinker & Prince, Lachter & Bever, Plunkett & Marchman, and Prasada & Pinker. Recently MacWhinney and Leinbach design a new ANN model in an attempt to answer some of the criticisms. They even pose a specific challenge claiming that no symbolic model can perform as well as their ANN. In this paper, we take up their challenge, and construct a Symbolic Pattern Associator (SPA) that can learn symbolic trees and rules from any input/output patterns. In an experimental comparison with the connectionist models based on the *same* set of verbs, the SPA is able to generalize much better from the training examples to unseen regular and irregular verbs. The SPA's results are also more psychologically realistic than ANN models when compared with human subjects. In addition, the SPA represents the acquired knowledge in symbolic rules which are meaningful to human observers. Such rules can be generalized further and can be integrated in other knowledge domains and language learning modules. Our results support the view that language learning and language processing is a rule-governed process.

ing the original results with a new connectionist model was made by MacWhinney and Leinbach [MacWhinney and Leinbach, 1991]. The connectionist models of learning the English past tense are specifically designed to refute the hypothesis that cognition in general, and language learning in particular, can best be explained as a rule-governed process where rules operate on symbolic representations. The debate about the validity of this Rules and Representation Hypothesis is central to the dispute between connectionism and the symbol-processing approach (cf. [Fodor and Pylyshyn, 1988; Pinker and Prince, 1988]). The proper way of debating is by contrasting competitive implementations. Thus, a symbolic implementation is needed that can be contrasted with the connectionist models. This is, in fact, a specific challenge posed by MacWhinney and Leinbach [MacWhinney and Leinbach, 1991], claiming that no symbolic methods would work as well as their model. In this paper we present a Symbolic Pattern Associator that can learn the past tenses of English verbs much better than the connectionist models. But in sharp contrast to the connectionist models, our SPA can express the acquired knowledge in production rules. Our results support the view that language learning and language processing is a rule-governed process.

2 Review of Previous Work

No previous symbolic methods have been tried on this task. We review the two main connectionist models of learning the past tenses of English verbs and the subsequent criticism in this section.

2.1 Rumelhart and McClelland's Model

Rumelhart and McClelland's PDP (Parallel Distributed Processing) learning model [Rumelhart and McClelland, 1986] is designed specifically to refute the Rules and Representations Hypothesis. They claim that the outcome of their experiment disconfirms the view that there exist explicit (inaccessible) rules that underlie human knowledge of language. Rumelhart and McClelland's model is based on a simple perceptron-based pattern associator interfaced with an input/output encoding/decoding network which allows the model to associate verb stems with their past tenses using a special Wickelphone/Wickelfeature phoneme-representation format. The learning algo-

1 Introduction

One of the most important connectionist contributions to language learning and language processing studies has been the attempt to model children's acquisition of the past tenses of English verbs. After the publication of the results of the first landmark connectionist simulation of this learning task [Rumelhart and McClelland, 1986], a number of critical reviews [Pinker and Prince, 1988; Lachter and Bever, 1988; Plunkett and Marchman, 1991; Pinker, 1991; Kim *et al.*, 1991; Prasada and Pinker, 1992] have appeared. A subsequent attempt at improv-

rithm is the classical perceptron convergence procedure. Rumelhart and McClelland's simulation use the "traditional" learning paradigm in which the training and the testing sets are *mutually disjoint*. The errors made by the model during the training process broadly follow the U-shaped learning curve in the stages of acquisition of the English past tense exhibited by young children. The testing sample consists of 86 "unseen" low frequency verbs (14 irregular and 72 regular) that are not randomly chosen. The *testing* sample results have a 93% error rate for the irregulars. The regulars fare better with a 33.3% error rate. Thus, the overall error rate for the whole testing sample is 43% – 37 wrong or ambiguous past tense forms out of 86 tested.

2.2 MacWhinney and Leinbach's Model

MacWhinney and Leinbach [MacWhinney and Leinbach, 1991] report a new connectionist model on the learning of the past tenses of English verbs. They claim that the results from the new simulation are far superior to Rumelhart and McClelland's results and that they can answer most of the criticisms addressed at the earlier model. The major departure from Rumelhart and McClelland's model is that the Wickelphone/Wickelfeature representational format is replaced with the UNIBET [MacWhinney, 1990] phoneme representational system which allows the assignment of a single-letter ASCII code to each phoneme. However, MacWhinney and Leinbach use special *templates* with which to code each phoneme and its position in a word. The actual input to the network is created by coding the individual phonemes as disjoint sets of phonetic features in a way similar to the coding of Wickelphones as Wickelfeatures. The network has two layers of 200 "hidden" units fully connected to adjacent layers. This number is arrived at through trial and error. In addition, MacWhinney and Leinbach's network has an additional special-purpose set of connections that allow it to *copy* the left-justified phonological form of the input *directly* onto the output in order to facilitate learning.

Altogether 2062 regular and irregular English verbs are selected for the experiment – 1650 of them are used for training (1532 regular and 118 irregular) but strangely only 13 *low frequency irregular verbs* are used for testing. (cf. [MacWhinney and Leinbach, 1991], p. 144). Training the network takes 24,000 epochs. At the end of training there are still 11 errors on the irregular pasts. MacWhinney and Leinbach believe that if they allow the network to run for several additional *days* and give it additional hidden unit resources, it probably can reach complete convergence. ([MacWhinney and Leinbach, 1991], p. 151) The error rate on the very small and biased test sample of 13 unseen irregular verbs is 69.2%! However, for no obvious reason they do not test their model on any of the unseen regular verbs. ([MacWhinney and Leinbach, 1991], p. 151) Further, more "hidden" features in representation that may facilitate learning (such as phonetic feature coding, templates, copy connections) are assumed than in the first ANN simulation by Rumelhart and McClelland. Thus, MacWhinney and Leinbach's new model actually represents a set-back

in comparison to Rumelhart and McClelland's model, rather than an improvement.

2.3 Criticism of the Connectionist Models

Previous and current criticism of the connectionist models of learning the past tenses of English verbs centre mainly on several issues [Pinker and Prince, 1988; Lachter and Bever, 1988; Plunkett and Marchman, 1991; Pinker, 1991; Kim *et al.*, 1991; Prasada and Pinker, 1992]:

- Very high error rates and low reliability of the experimental results.
- Inappropriateness of the training and testing procedures.
- "Hidden" features of the representational format and the network architecture that facilitate learning.
- Opaque knowledge representation and integration of acquired knowledge.

3 The Symbolic Pattern Associator

Unfortunately, in the absence of a symbolic learning model that could match even the modest results achieved by Rumelhart and McClelland's and MacWhinney and Leinbach's models, the connectionists could set aside all criticism of their models. Noting the absence of rival symbolic models MacWhinney and Leinbach pose a specific challenge to the critics of connectionism:

If there were some other approach that provided an even more accurate characterization of the learning process, we might still be forced to reject the connectionist approach, despite its successes. The proper way of debating conceptualizations is by contrasting competitive implementations. To do this in the present case, we would need a symbolic implementation that could be contrasted with the current implementation. ([MacWhinney and Leinbach, 1991], p.153)

We take up MacWhinney and Leinbach's challenge and present a symbolic learning model that can learn the past tenses of English verbs much better than any PDP model so far. We hope that the success of our symbolic learning model will demonstrate the vacuousness of the connectionist criticism of the Rules and Representations Hypothesis.

The core of our model is a general purpose Symbolic Pattern Associator (SPA for short). The SPA is a very general and efficient symbolic algorithm for associating arbitrary patterns which have a finite number of attributes. If all the attribute values are discrete or symbolic, ANNs do not have any systematic advantage over the SPA in their prediction or generalization ability. On the contrary, as this paper will demonstrate exactly the opposite is true in the task of learning the English verb past tenses.

3.1 The Architecture of the Symbolic Pattern Associator

The SPA is based on c4.5 [Quinlan, 1990] which is an implementation of the ID3 learning algorithm [Quinlan, 1986]. ID3 is a program for inducing classification rules in the form of decision trees from a set of given examples which may contain unknown or noisy entries. It uses information gain as a criterion to select attributes as roots of the subtrees. The divide and conquer strategy is recursively applied in building subtrees, until all remaining examples in the training set belong to a single concept (class); then a leaf is labeled as that concept. The information gain guides a one-step look-ahead greedy heuristic search for the *most relevant* or *discriminative* attribute that maximally reduces the entropy in the divided set of the examples. The use of this heuristic results in building *small* decision trees which have a good generalization power in general.

Although ID3 has been very successful in competing with the learning and representational capabilities of neural networks, they are essentially limited in the type of learning task that they can perform. If the task is to learn to classify a set of different patterns into several *mutually exclusive categories*, as many of the most successful connectionist simulations actually do, ID3 has been shown to perform as well as, or even sometimes better than neural networks [Shavlik *et al.*, 1991; Atlas *et al.*, 1990]. However, if the task is to classify a set of (input) patterns into (output) patterns of many attributes, ID3 cannot be applied directly. The reason is that since ID3 treats the different output patterns as mutually exclusive classes, the number of classes would be exponentially large, and more importantly, any generalization among attributes within the patterns would be lost.

To turn ID3 or any similar classification system into general purpose symbolic pattern associators, we apply ID3 on all output attributes and combine individual decision trees into a "forest" or set of trees. Each tree takes as input the set of all attributes in the input patterns, but is concerned with determining the value of one attribute in its output pattern. More specifically, if the attributes of input (ι_1 to ι_n) and output (ω_1 to ω_m) patterns can be represented as:

$$\iota_1, \dots, \iota_n; \omega_1, \dots, \omega_m$$

then a total of m decision trees are built, one for each output attribute ω_i ($1 \leq i \leq m$) taking all input attributes $\iota_1, \dots, \iota_n$ per tree. Once all of m trees are built using ID3, they can be used jointly to determine the output pattern $\omega_1, \dots, \omega_m$ from any input pattern $\iota_1, \dots, \iota_n$. Clearly, any pattern association task can be solved by the SPA. Similar to ID3, our SPA is very general and simple to use: one only needs to prepare input and output attribute name files, and training and testing files containing pairs of patterns.

An important feature of the SPA is explicit knowledge representation. Since decision trees can be transformed into propositional rules, sets of rules for all output attributes can be obtained. These rules are sometimes meaningful to the human observer. In addition, further processing and integration of these rules with high-level

knowledge is possible (see SECTION 4.4).

3.2 Application of the SPA to the Learning of Past Tenses

In order to answer the challenge posed to the symbolic processing approach by the connectionist models of learning the past tenses of English verbs, we structured our experiment in such a way as to guarantee a maximal common basis for a comparison between the performance of these models and the SPA. Since MacWhinney and Leinbach used a much larger set of verbs than Rumelhart and McClelland, we borrowed their original list of verbs. The set that we use contains 1404 stem/past tense pairs. Because of the large overlap in the verb sets used in our experiment and in MacWhinney and Leinbach's simulation, our results can be compared directly with their results. Even though Rumelhart and McClelland used altogether only 506 verbs, the overlap between the two sets is still sufficient for a similar comparison.

However, in order to guarantee a more thorough testing of the SPA and a fair comparison with the connectionist models, we decided on several significant changes in MacWhinney and Leinbach's, and Rumelhart and McClelland's experimental set up:

- We do not use Wickelphones/Wickelfeature, phonetic features, or consonant/vowel templates. Instead, we use simple and straightforward left-to-right ASCII coding of phonemes in UNIBET [MacWhinney, 1990] format. Also, unlike MacWhinney and Leinbach's model, we do not include any special purpose copying devices that would allow for a present tense form of a verb to influence directly the output of our system. Below are some input/output pairs of verbs (shown in the spelling form) (abandon - abandoned), (accelerate - accelerated), (arise - arose), and (become - became) used in SPA (where \$ represents blank):

6,b,&,n,d,6,n,\$,\$,\$,\$... > 6,b,&,n,d,6,n,d,\$,\$,\$,\$...
I,k,s,E,1,6,r,e,t,\$,\$,\$... > I,k,s,E,1,6,r,e,t,I,d,\$,\$,\$...
6,r,3,z,\$,\$,\$,\$,\$,\$,\$... > 6,r,o,z,\$,\$,\$,\$,\$,\$,\$...
b,I,k,6,m,\$,\$,\$,\$,\$,\$,\$... > b,I,k,e,m,\$,\$,\$,\$,\$,\$,\$...

- To guarantee unbiased and reliable results, we use only *randomly drawn* training and testing samples in several independent trials. This allows us, in contrast to the connectionist simulations, to vary the sizes of the training and testing samples at different trials and thus to achieve a more robust estimate of the inductive generalization capabilities of our model.
- Two methods of random sampling are implemented.
(1) Adopting the traditional learning paradigm, both training and testing samples contain at most one copy of each stem/past-tense verb pairs. Both training and testing samples are drawn randomly from the whole set of 1404 verb pairs with *no* overlap between training and testing. Both regular and irregular verbs may be drawn randomly into training/testing samples. This set-up best tests the generalization power of the SPA.

(2) Adopting the PAC [Valiant, 1984] learning paradigm, *both* training and testing samples are drawn randomly from the same set of verbs (with replacement) according to a probability distribution proportional to the Kucera-Francis frequency of the verbs [Kucera and Francis, 1967]. Thus, the error rate in the testing sample converges to the probability of incorrect response of a verb stem randomly drawn according to the same probability distribution. Verb pairs may have multiple copies in training/testing samples, and training and testing samples may overlap. This set-up best reflects the actual learning of children who encounter frequently-used verbs more often. Using testing and training samples with possible overlap best reflects how well they could have learned in the same environment.

Note that no information about regular or irregular forms is provided in both training and testing input/output verb pairs. It is SPA's task to learn rules for regular and irregular verbs, and to apply acquired rules to unseen verbs, which can be either regular or irregular, unknown to SPA.

4 Learning Results of the SPA

4.1 Performance Comparison between the SPA and the Connectionist Models

The results from the two series of experiments allow us to make a direct comparison between the inductive generalization capabilities of the SPA and the neural networks used in the connectionist simulations. TABLE 1 compares the error rates of Rumelhart and McClelland's model (R&M), MacWhinney and Leinbach's model (M&L) and the two averaged results (both with 1,000 training examples) achieved by the SPA by adopting the "traditional" and the PAC learning paradigms¹.

	R&M	M&L	SPA Traditional	SPA PAC
Templates	Yes	Yes	No	No
Training Sample Size	420	1650	1000	1000
Test Sample Size	86	13	300	500
Random Samples	No	No	Yes	Yes
Results Averaged Over Multiple Trials	No	No	Yes	Yes
% Correct On Training				
Total	97.0	99.3	100.0	100.0
Regulars	98.0	100.0	100.0	100.0
Irregulars	95.0	90.7	100.0	100.0
% Correct On Testing				
Total	57.0	N/A	82.2 ± 1.6	93.9 ± 0.9
Regulars	66.7	N/A (!)	89.8 ± 0.9	92.7 ± 0.8
Irregulars	7.1	30.8*	17.6 ± 8.2	95.4 ± 1.2
Avg. Learning Time	280 hrs	weeks(?)	70.2 sec	59.5 sec

* This number represents a single test on 13 irregular verbs.

Table 1: A comparison between SPA and the Connectionist Models

It is indisputable that the SPA achieves a much better overall performance than the connectionist models. Although the training results of the three models are similar, the crucial measure of the success of a learning model – the *testing* results – are clearly in favour of the symbolic

¹The average learning time for the SPA is the CPU time obtained from the MIPS M/2000, which is rated at 20 MIPS.

model. If we adopt the traditional learning paradigm, the overall success rate of the SPA is well above the unrealistically low success rate of Rumelhart and McClelland's model. MacWhinney and Leinbach did not test the overall inductive capabilities of their model at all. The result of testing irregular verbs is not reliable at all: it based on one test of 13 specifically chosen irregular verbs. On the regular verbs, our model shows a much stronger tendency to capture the regularities underlying regular past formation than do the connectionist models. However, the results of the three models on the irregular pasts are not psychologically realistic. One possible solution is to adopt the PAC learning paradigm which resembles much closer the actual learning conditions for humans and can allow a proper use of frequency information. We cannot speculate whether the connectionist models would achieve better results adopting the PAC learning paradigm but it is clear that the performance of the SPA, especially on the irregulars, improves dramatically.

Another disadvantage for the connectionist models is that they are extremely slow. Possibly, because of this the connectionist simulations did not allow the training/testing of several different randomly drawn samples of verbs. This is quite unfortunate because there are hidden pitfalls in carrying out just a single test on a single verb set; the results, especially on irregular verbs, tend significantly to alternate from sample to sample.

It is interesting to study the kind of errors that the SPA makes using the traditional learning paradigm (no overlap between training and testing). Incorrect results for irregular verbs are mainly due to the fact that they are treated by the SPA as regular (Reg.), but the no-change (No-chg.), vowel change (V-chg.), and consonant change (C-chg.) errors contribute relatively little to the overall error rate. Note that most of the no-change, vowel change and consonant change are more or less "meaningful", since children may make the same mistakes (See Section 4.2). In contrast to the connectionist simulations, very few errors contain phonetically unrealistic transformations (Imp.). TABLE 2 summarizes the types of errors.

Verb Type	Sample Size	Correct (%)	Incorrect (%)			
			Reg.	No-chg.	V-C-chg.	Imp.
Irregular	35	25.7	51.4	8.6	11.4	2.9
Regular	265	89.1	n/a	6.4	3.78	0.75
Total	300	81.7	6.0	6.7	4.63	1.0

Table 2: Types of Responses

4.2 Comparison to Human Performance on Pseudo-Verbs

It is difficult to compare SPA (or ANNs) performance with human adults in predicting unseen verbs simply because all irregular verbs are frequently used in everyday life (in fact, the 10 most frequently used English verbs are irregular). Thus adults know most of the irregular verbs already. In order to achieve direct comparison with human subjects on "novel" verbs, we perform an experiment using Prasada and Pinker (1992) *pseudo-verb*

testing paradigm to see if SPA produces psychologically realistic results.

Prasada and Pinker create 60 English-like pseudo-verbs. The verbs are divided into two main groups – irregular-like and regular-like pseudo-verbs. Each of these groups are subdivided into three classes – Prototypical, Intermediate, and Distant [Prasada and Pinker, 1992], which resemble to different degrees ordinary irregular and regular English verbs. Prasada and Pinker carry out a test of Rumelhart and McClelland’s model using the pseudo-verb paradigm. They train the network on the 422 verbs (341 regular and 81 irregular) which are used in Rumelhart and McClelland’s simulation. The model is trained for 200 epochs and then is tested on the pseudo-irregular verbs. To assess the psychological validity of these results, Prasada and Pinker carry out an experiment with 24 native English speakers. They want to test specifically the likelihood of people actually producing regular and irregular past tense forms, so that it would be possible to compare directly the human results with the connectionist simulations. As we can see (also on FIGURE 1 (a) and (b)), the difference between Rumelhart and McClelland’s simulation and human subjects’ responses is quite large.

In order to test the inductive capabilities of the SPA in comparison with human subjects and with Rumelhart and McClelland’s connectionist model we perform the following experiment: We train the SPA on all 1,404 of the verbs in the original verb set. The training set includes 1,262 regular and 142 irregular English verbs. When we test the SPA on the 60 pseudo-verbs, all of the past tense forms produced by our SPA are “good” pseudo-past forms. That is, in contrast to the Rumelhart and McClelland’s model, no normal English verbs and no bizarre consonant changes or strange suffixes are produced.

We list (in the spelling form) some of Prasada and Pinker’s pseudo-verbs and the past tense forms produced by the SPA after it learns from 1404 (real) verbs. As we can see, SPA is making “intelligent” decisions in predicting past tenses of the pseudo-verbs: some are treated as irregular and SPA produces sensible irregular past tenses, while others are treated as regular verbs.

spling → splang
 skring → skrung
 sprink → sprink
 queed → queeded (SPA classifies it as a regular verb)
 goav → wev (Why? SPA has learnt go → went!)

On FIGURE 1 (a) and (b), our results for the pseudo-irregular and for the pseudo-regular pasts are plotted together with the results from the human subjects and from the Rumelhart and McClelland’s model. Our results for the pseudo-irregular class are actually closer to the result from the human subjects than the results of Rumelhart and McClelland’s model. Especially for the pseudo-regular items the SPA performed much closer to the human subjects than the connectionist model. This demonstrates that the SPA’s performance on both pseudo-irregulars and pseudo-regulars is psychologically realistic, whereas the connectionist model fails at the pseudo-regulars and is not very close to the human sub-

jects on pseudo-irregulars.

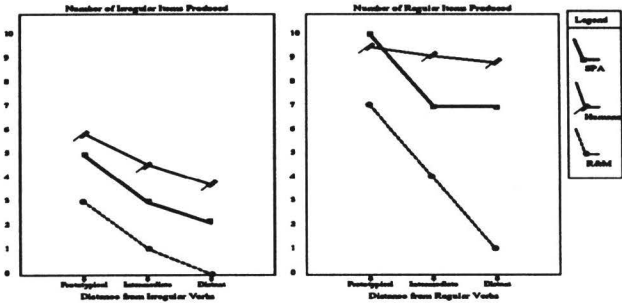


Figure 1: (a) Pseudo-Irregular (b) Pseudo-Regular

4.3 U-shaped Learning Curve of Regular and Irregular Verbs

A U-shaped learning curve has been observed in the acquisition of regular and irregular verbs by young children. Very briefly, during the *early* phase in the acquisition of the English past tense, children apparently learn a relatively small number of high frequency irregular and regular verbs correctly. During the *intermediate* phase, children learn an increased number of medium frequency verbs, the majority of which are regular. Children begin to overregularize the irregular verbs, even the ones that they have correctly learned before. In the *late* phase, competence similar to that of mature speakers who produce both regular and irregular past tenses correctly is observed.

The SPA is also able to model the U-shaped curve apparently exhibited by children in the early stages of development that Rumelhart and McClelland try to model [Rumelhart and McClelland, 1986]. To model the U-shape using the SPA, we change the training size and m , a parameter in the program c4.5 (ID3). The parameter m represents the minimal number of exceptions that c4.5 can tolerate without building new branches to accommodate these exceptions. In all of our other experiments we report in the paper, m is fixed at 1 (since there is no noise in our data). Four stages of successive past-tense acquisition are created. Training and testing samples are drawn randomly adopting the PAC learning paradigm. This paradigm best reflects the natural learning environment in which learning and testing are based on the same probability distribution. The stages are as follows:

1. *Training size = 500 and $m = 1$.* The verb set is small and the SPA learns/memorizes 100% correctly the training sample. This models the *early phase* in children’s acquisition of the English past tense. Since some irregular verbs have very high frequency, they are reproduced quite well in the testing sample by simply memorizing them from the training sample.
2. *Training size = 1000 and $m = 5$.* When more and more regular and irregular verbs are presented to the SPA, m can be set to a larger number to al-

low generalization of regular verbs. Since the total of regular verbs out-numbers the total of irregular verbs in the training sample, such generalization leads to only a small increase in the error rate on regular verbs.

3. Training size = 1000 and $m = 10$. SPA generalizes further. At this time, most learned irregular verbs are treated as "noise" by the SPA and they are simply regularized. The second and the third stages model the *intermediate phase* in children's acquisition of the English past tense.
4. Training size = 3000 and $m = 10$. m is now fixed but the training sample is increased. The number of appearances of irregular verbs is increased, and when it is greater than m , these irregular verbs will not be treated as "noise" and branches in the trees for these irregulars are built up. At this stage, the SPA captures rules for regulars and reproduces irregulars both very well. This models the *late phase* in children's acquisition of the English past tense, where competence of mature speakers is observed.

See FIGURE 2 for the U-shaped curve produced by the SPA.

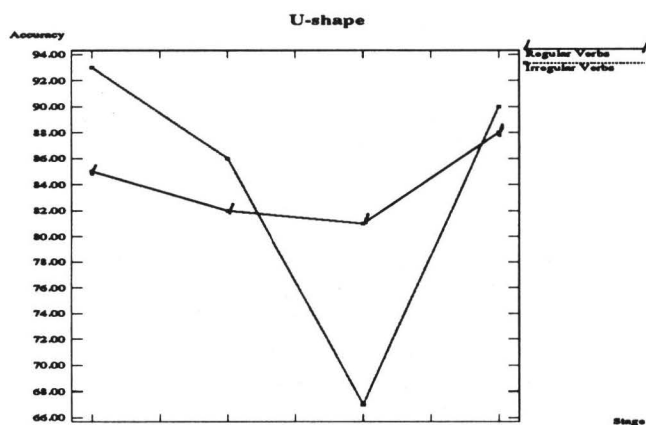


Figure 2: U-shaped curve of learning regular/irregular verbs.

4.4 Representation of Acquired Knowledge and Higher-Level Processing

We consider the comparative results previously presented to be a convincing proof that the SPA outperforms both MacWhinney and Leinbach's, and Rumelhart and McClelland's connectionist models in the task of learning the past tenses of English verbs. However, symbolic learning systems such as our SPA have an additional significant advantage over the connectionist learning models. In sharp contrast to the complete absence of knowledge representation capacities in connectionist pattern associators, the SPA can express the acquired knowledge in the form of *production rules*; this is an explicit knowledge representation format that opens the possibility for these rules to be generalized and integrated with lexical, morphological, and syntactic rules. Thus, unlike the connectionist models, the SPA does not

leave the integration of the acquired linguistic knowledge as a complete mystery.

The decision trees produced during the learning process are a useful symbolic format that lends itself to an easy conversion into production rules. In this particular experiment, the production rules resulting from the conversion of all trees can be classified into three different groups, corresponding to the three different types of branches of a "typical" tree – identity-mapping rules, suffix-adding rules, and exceptional changes rules. TABLE 3 presents a sample of the different types of rules produced by the SPA in one of our experiments. (Note that the indexes have the following meanings: ι_n gives the value of the n^{th} component of the input vector; ω_n gives the value of the n^{th} component of the output vector, and the rule is derived from the n^{th} decision tree.)

Identity Rules	If $\iota_1 = p$, then $\omega_1 = p$ If $\iota_2 = p$, then $\omega_2 = p$ If $\iota_3 = d$, then $\omega_3 = d$ If $\iota_4 = p$, then $\omega_4 = p$ If $\iota_5 = i$, then $\omega_5 = i$
Exceptional rules (for $go \mapsto went$)	If $\iota_1 = g$ and If $\iota_2 = o$, then $\omega_1 = w$
Suffix-adding rules	If $\iota_4 = p$ and $\iota_5 = \$$, then $\omega_5 = t$ If $\iota_4 = r$ and $\iota_5 = \$$, then $\omega_5 = d$ If $\iota_5 = p$ and $\iota_6 = \$$, then $\omega_6 = t$ If $\iota_5 = r$ and $\iota_6 = \$$, then $\omega_6 = d$ If $\iota_6 = p$ and $\iota_7 = \$$, then $\omega_7 = t$ If $\iota_6 = r$ and $\iota_7 = \$$, then $\omega_7 = d$

Table 3: Decision tree rules

If we look at the first of the suffix-adding rules we can see that it expresses the following information: look at the fourth phoneme of an input verb form and look at the fifth, if the fourth is /p/ and the fifth is blank, i.e. if we are at a verb ending, then add the unvoiced /t/ as the fifth phoneme to the output form of the verb. The second rule is: if the fourth input phoneme is /r/ and the fifth is blank \$, then add the voiced /d/ as the fifth phoneme to the output form of the verb, etc. Since all trees that map input patterns onto the middle portion of the output patterns have a very similar branch structure, it is not surprising to find that they all produce very similar production rules. For example, in TABLE 3 we have followed all the productions for the phonemes /r/ and /p/ produced by the fifth, sixth, and seventh trees. In fact, most trees for ω_i ($i \geq 4$) have a very similar structure. This suggests that the SPA has learned to distinguish between verbs that receive a voiced or unvoiced suffix 'ed' in the regular past tense formation.

SPA itself cannot generalize from the phonetic rules for verbs of particular length to *general* phonetic rules for verbs of any length. In essence, SPA (and ANNs) cannot make the step from *propositional* production rules to *first-order* rules. But because, in contrast to the neural networks, the SPA is able to represent its position-dependent knowledge in explicit rule form, it is not difficult to see how this knowledge can be generalized to first-order form. First, the propositional rules derived

by the SPA are changed to a form in which the position indexes are represented as terms. Then there are a number of symbolic learning programs, such as GOLEM [Muggleton and Feng, 1990], which utilize the *least general generalization* (LGG) [Plotkin, 1970] that can be applied to produce automatically new rules that are position independent first order phonetic rules like:

For all n , if $\iota(n) = k$ & $\iota(n+1) = \$$ then $\omega(n+1) = t$ where $\iota(n)$ stands for *any* phoneme at position n . Clearly, this is a rule in first-order logic and it applies to verbs of any length. Thus, adding suffix to "verb stem" is represented by this rule.

The rules that we have induced by generalizing the position relevant phonetic rules are still purely phonetic. We can easily see that they will give the wrong answer if the phonemes over which they range are part of an adjective or a noun instead of a verb stem. Clearly, we need to integrate these phonetic rules with *morphological* and *lexical* information to produce the correct rules for the past-tense formation in English. If purely phonetic information cannot be represented in a form that can be readily interfaced and integrated with lexical, morphological, and syntactical information available to other language processing subsystems, it is hard to imagine how adult linguistic competence is developed. The connectionist models consider the acquisition of the English past tense as a totally isolated affair and fail to explain how the knowledge acquired by the networks is integrated with other types of linguistic knowledge. Symbolic approaches, on the contrary, can represent the acquired knowledge as sets of production rules that can easily be unified into more general first-order phonetic production rules. This highlights the great advantages in terms of representation of acquired knowledge that the SPA has over the connectionist models of language learning.

5 Conclusions

In the paper we present a symbolic method for learning the past tenses of English verbs. In comparison to the connectionist approaches, our method has several distinctive advantages, and is much closer to actual human performance. Our results support the view that language learning and language processing is a rule-governed process.

Acknowledgement

We thank B. MacWhinney for providing us with the verb data used in their simulation.

References

- [Atlas *et al.*, 1990] L. Atlas, R. Cole, J. Connor, M. El-Sharkawi, R. Marks, Y Muthasamy, and E. Barnard. Performance comparisons between backpropagation networks and classification trees on three real-world applications. In D. Touretzky, editor, *Advances in Neural Information Processing Systems, Vol 2*, pages 622 – 629. San Mateo, CA: Morgan Kaufmann Inc., 1990.
- [Fodor and Pylyshyn, 1988] J.A. Fodor and Z. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. In S. Pinker and J. Mehler, editors, *Connections and Symbols*, pages 3 – 71. Cambridge, MA: MIT Press, 1988.
- [Kim *et al.*, 1991] J. Kim, A. Pinker, A. Prince, and S. Prasada. Why no mere mortal has ever flown out to center field. *Cognitive Science*, 15:173 – 218, 1991.
- [Kucera and Francis, 1967] H. Kucera and W. Francis. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press, 1967.
- [Lachter and Bever, 1988] J. Lachter and T. Bever. The relation between linguistic structure and associative theories of language learning – a constructive critique of some connectionist learning models. In S. Pinker and J. Mehler, editors, *Connections and Symbols*, pages 195 – 247. Cambridge, MA: MIT Press, 1988.
- [MacWhinney and Leinbach, 1991] B. MacWhinney and J. Leinbach. Implementations are not conceptualizations: Revising the verb model. *Cognition*, 40:121 – 157, 1991.
- [MacWhinney, 1990] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Erlbaum, 1990.
- [Muggleton and Feng, 1990] S. Muggleton and C. Feng. Efficient induction of logic programs. In *Proceedings of first International Conference on Algorithmic Learning Theory*. OHMSHA Tokyo, 1990.
- [Pinker and Prince, 1988] S. Pinker and A. Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. In S. Pinker and J. Mehler, editors, *Connections and Symbols*, pages 73 – 193. Cambridge, MA: MIT Press, 1988.
- [Pinker, 1991] A. Pinker. Rules of language. *Science*, 253:530 – 535, 1991.
- [Plotkin, 1970] G.D. Plotkin. A note on inductive generalization. In B. Meltzer and D. Michie, editors, *Machine Intelligence 5*, pages 153–163. Elsevier North-Holland, 1970.
- [Plunkett and Marchman, 1991] K. Plunkett and V. Marchman. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 38:43 – 102, 1991.
- [Prasada and Pinker, 1992] S. Prasada and A. Pinker. Generalization of regular and irregular morphological patterns. In press, *Language and Cognitive Processes*, 1992.
- [Quinlan, 1986] R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81 – 106, 1986.
- [Quinlan, 1990] J.R. Quinlan. Probabilistic decision tree. In Y. Kodratoff and R.S. Michalski, editors, *Machine learning: An artificial intelligence approach*, volume III. Morgan Kaufmann Publishers, 1990.
- [Rumelhart and McClelland, 1986] D. Rumelhart and J. McClelland. On learning the past tenses of english verbs. In D. Rumelhart, J. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing Vol 2*, pages 216 – 271. Cambridge, MA: MIT Press, 1986.
- [Shavlik *et al.*, 1991] J.W. Shavlik, R.J. Mooney, and G.G. Towell. Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning*, 6(2):111 – 144, 1991.
- [Valiant, 1984] L.G. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.