Connectionist Rules of Language

Gert Westermann
Department of Computer Science
Technical University of Braunschweig
38106 Braunschweig, Germany
westerma@ibr.cs.tu-bs.de

Rainer Goebel
Max Planck Institute for Brain Research
Deutschordenstr. 46
60528 Frankfurt, Germany
goebel@mpih-frankfurt.mpg.d400.de

Abstract

A modular connectionist network is described that learns the German verb paradigm. The architecture of the network is in accordance with the rule–associative memory hypothesis proposed by Pinker (1991): it is composed of a connectionist short-term memory enabling it to process symbolic rules and an associative memory acting as a lexicon. The network successfully learns the German verb paradigm and generalizes to novel verbs in ways that correspond to empirical data. Lesioning the model gives further evidence for the rule–associative memory hypothesis: When the lexicon is cut off, the network strongly overgeneralizes the regular participle, indicating that regular forms are produced with the short-term memory but irregular forms rely on the lexicon. However, in contrast to the rule–association theory, the two paths are not strongly dissociated, but both the short-term memory and the lexicon work together in producing many participles. The success of the network model is seen as evidence that emergent linguistic rules need not be implemented as rules in the brain.

Introduction

It has recently been argued that certain parts of the language system are represented in the brain by a dualistic framework combining rule manipulation with associative memory (Pinker, 1991). For example, in the English past tense regular forms are claimed to be produced by a symbolic rule ("*add -ed to the verb stem*") while irregular verbs are stored in an associative memory together with links to their past tense forms. When a past tense form is to be produced, first the associative memory is searched for a matching entry, and only if none is found the *default* rule is applied (Marcus et al., 1993). The cases in which the default rule applies are called *regular*. Regularity is independent from frequency, because the default rule might apply only to a minority of all cases. Marcus et al. (1993) argued that this is true for both the German noun plural and the German participle: The default noun plural ending *-s* has a type frequency of less than 8% and a token frequency of less than 2% of all nouns. The regular verb participle accounts for 45% of all verb types and only 17% of all verb tokens. This discrepancy between regularity and frequency is in contrast with the English past tense, where 86% of all verb types (40% of all tokens) are regular.

The instances in which the regular cases do not form the majority of all cases can be seen as a touchstone for connectionist models. This is because homogenous connectionist models (e.g. MacWhinney and Leinbach, 1991, Rumelhart and McClelland, 1987) learn by exploiting the statistical regularities of the input data and therefore necessarily assign the regular status to the most frequently occuring input.

In this paper, a modular connectionist network is described that implements the rule–associative memory framework for the German verb paradigm. The network consists of a connectionist short-term memory enabling it for symbol manipulation (see Goebel, 1991) and a lexicon in the form of an associative memory. It is shown that the network learns the verb inflection paradigm and generalizes to novel verbs in a way comparable to empirical data. The internal representations developed by the network follow the rule–associative memory theory (Pinker, 1991) in that most regular forms are produced solely with the short-term store while the irregular forms are handled by the lexicon. It is demonstrated, however, that in contrast to this theory the rule path and the associative memory are not strongly dissociated from one another, but they interact in producing the correct output for certain verbs. The fact that the network model displays ruke-like behavior without hard-wired rules suggests that emergent linguistic regularities need not be implemented as explicit rules in the brain.

Below, the German verb paradigm is first briefly reviewed. Our experiments, including the data and the network architecture, are then described, followed by a detailed analysis of the network's performance and of its internal representations. Finally, we discuss the implications to the rule–associative memory theory (Pinker, 1991).

The German Verb Paradigm

The German verb paradigm is illustrated in table 1 (for a more detailed account, see Marcus et al., 1993). Each verb has three paradigmatic forms: an infinitive, a preterite, and a participle. *Infinitives* are formed by adding the suffix `-en` to the verb stem: For example, the verb `spielen` has the stem `spiel` and the infinitive is formed by adding `-en` . *Preterites* are formed in different ways,

Table 1: The structure of the German verb paradigm

Infinitive	Preterite	Participle
Weak verbs		
spielen (to play)	spielte (played)	gespielt (have played)
wollen (to want)	wollte (wanted)	gewollt (have wanted)
Strong verbs		
kommen (to come)	kam (came)	gekommen (have come)
gehen (to go)	ging (went)	gegangen (have gone)
Mixed verbs		
können (can)	konnte (could)	gekonnt (have been able to)
denken (to think)	dachte (thought)	gedacht (have thought)

but they are not common in spoken language and hence play no role in our simulation experiments. *Participles* are formed by adding a suffix (either `-t` or `-en`) to the stem which might be different from the infinitive stem. Participles whose stems have primary stress on the first syllable (most of them do) receive the prefix `ge-`.

There are three verb classes, weak verbs, strong verbs, and mixed verbs, which differ in the ways that preterites and participles are formed: The participle of weak verbs is always formed by adding the `-t` suffix and the `ge-` prefix (where applicable) to the unchanged stem: `ge- spiel -t` . The participle of strong verbs is formed by adding the suffix `-en` (and the prefix `ge-`) to the participle stem which may or may not be different from the infinitive stem: For example, `kommen` does not change its stem: `ge- komm -en` , while `gehen` does: `ge- gang -en` . Mixed verbs change their stems like strong verbs, but they receive the suffix `-t` like weak verbs: `können` – `ge- konn -t` . Only one verb has completely idiosyncratic preterite and participle forms: `sein` (to be) – `war` – `gewesen`.

Marcus et al. (1993) give evidence that the weak verbs constitute the regular (default) case for the German verb paradigm.

Experiments

The goals of the experiments were threefold: First, to examine whether the connectionist model could learn the task of acquiring the German verb paradigm which is claimed to employ symbolic rules. Second, to test the network's ability to generalize to novel verbs and compare the results with data from psycholinguistic experiments. Third, to analyze the internal representations developed by the network in learning the German verb paradigm and relate them to the dualistic framework proposed by Pinker (1991).

Data

The data for the experiments was taken from a corpus of spoken utterances by German children in the first grade of elemantary school (Pregel and Rickheit, 1987). This corpus consisted of 824 types with 7,468 tokens. Since it contained a great number of composite verbs, all separable prefixes were removed because they do not influence formation of the participle. From the resulting corpus 100 types were randomly extracted, yielding 944 tokens. Since it was not given in this corpus, the ratio of the infinitive to participle forms of each verb was taken from the CELEX database. The tokens of each verb in the data were then divided according to this ratio. When a token did not appear in CELEX it was given the frequency 1 for both the infinitive and the participle form, thus guaranteeing the appearence of each verb at least once in each form.

For the experiments half the number of tokens for each verb form were used. The final data set thus contained 100 types with 538 tokens, of which 388 were in the infinitive form and 150 in the participle form.

An analysis of this data set yielded the following structure:

Verb class		
weak (regular):	52% types,	45.33% participle tokens
strong (irregular):	41% types,	50.00% participle tokens
mixed:	7% types,	4.67% participle tokens

Although weak verbs had a slight majority counting the types, they only accounted for 45.33% of the participle tokens. Even together with the mixed verbs (which also have the -t ending) they did not present a majority of the participle tokens. However, a comparison of this corpus with one for adult language (Ruoff, 1981) showed significant differences: In the adult corpus, only 45% of the types and 17% of the tokens belonged to the weak class in contrast to 52% and 45.33% for the child data, respectively. The vocabulary of children seems to comprise a higher ratio of weak verbs than that of adults, and these weak verbs seem to be used more often. In fact Pregel and Rickheit (1987) remarked that e.g. the frequency of the weak verb `haben` is 10.18% of all verb usages for children but only 4.65% for adults. The vocabulary of children is much less diverse than that of adults and it is obviously more slanted towards weak verbs. Whether this difference has implications for language acquisition

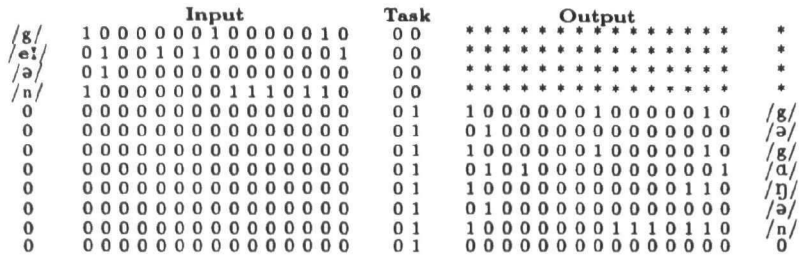

Figure 1: A typical input-output sequence for the network. The input is the verb **gehen**, the output the participle **gegangen**. The stars ('*') mark "don't care" states in which the activation of the units is irrelevant.

needs to be further investigated.

For the input to the network model the verbs were transcribed to phonetic writing and then encoded with phonetic featurevectors. Each of 56 phonemes was represented by a string of 14 features following Wurzel (1981). An additional feature was added to indicate whether a vowel was stressed. This was necessary to determine whether the participle would receive the **ge-** prefix (see above). In the feature vectors, '1' indicated the presence of a particular feature and '0' its absence.

Task

The task to be learned by the network model was as follows: A verb was presented to the input layer as a sequence of phonetic feature vectors. After presenting the input, one of the two task units was activated to determine the required output which was either a phonetic sequence for the infinitive form (that is, a duplication of the previous input) or the participle form of the verb. Figure 1 shows a typical sequence in which the input is the word **gehen** and the output is the participle of **gehen**, that is, **gegangen**.

Network Architecture

The network (see figure 2) was composed of two subsystems following the dualistic framework (see also Indefrey and Goebel, 1993): a short-term memory and a phonological lexicon. The short-term memory had fast weights enabling it to store a sequence of feature vectors after one presentation and retain the stored information (for a detailed description of the architecture see Goebel 1990, 1991). As it has been argued in Goebel (1991), short-term memory and selective attention are prerequisites for the ability to manipulate symbols. The short-term store thus represented the rule path of the dualistic framework.

The task of the phonological lexicon was to act as an associative memory for the input verbs transforming the input sequence of phonemes to a localist representation which could then act as a "plan vector" for determining the output. The lexicon consisted of two layered self-organizing feature maps. The lower map, which received the phonetic feature vectors from the input layer, was organized in a 12x12-units plane. It had been pre-trained so that all 56 different phonemes were laid out over the map. Each unit was connected to itself with a weight of 0.9. This allowed for a slow decay of unit activation after a phoneme had been presented. Thus, when the phoneme sequence of a verb was presented it left a unique trace on the map.

The phoneme map was used as input for the upper self-organizing feature map. This feature map had 400 units. It was trained so that each trace left by a verb on the phoneme map would be clustered to a single maximally responding unit, roughly corresponding to an entry in the phonological mental lexicon.

Out of the 100 different verbs, 88 clustered to distinct units. The remaining twelve verbs were clustered in pairs of two. Generally, rhyming verbs were grouped together on the map, for example, **heißen**, **reißen**, **beißen**, and **schmeißen** clustered to four neighboring units.

The lexicon had connections to the short-term store via a control unit. The idea behind these connections was that when the participle form was to be produced, the lexicon could inhibit activation of the short-term store until the prefix **ge-** had been produced. After **ge-** had been put out, the short-term store would be activated by the lexicon and replay the sequence for the input verb. Together with the plan vector from the lexicon, the correct participle stem and ending could then be formed.

The network model operated in two stages, a recognition stage and a production stage. In the recognition stage, the input sequence was presented and routed to the two subsystems, short-term memory and lexicon. After training of this recognition stage, the lexicon locally represented the input verb while the short-term memory held the sequence of phonemes. In the production stage, the activated lexical entry together with the sequentially accessible short-term store produced the required output sequence.

Performance

The production stage of the network was trained on the corpus of 538 tokens for 400 epochs with the back-propagation through time algorithm (Rumelhart et al., 1986). Throughout the training process, the performance of the network was tested at intermediate stages.

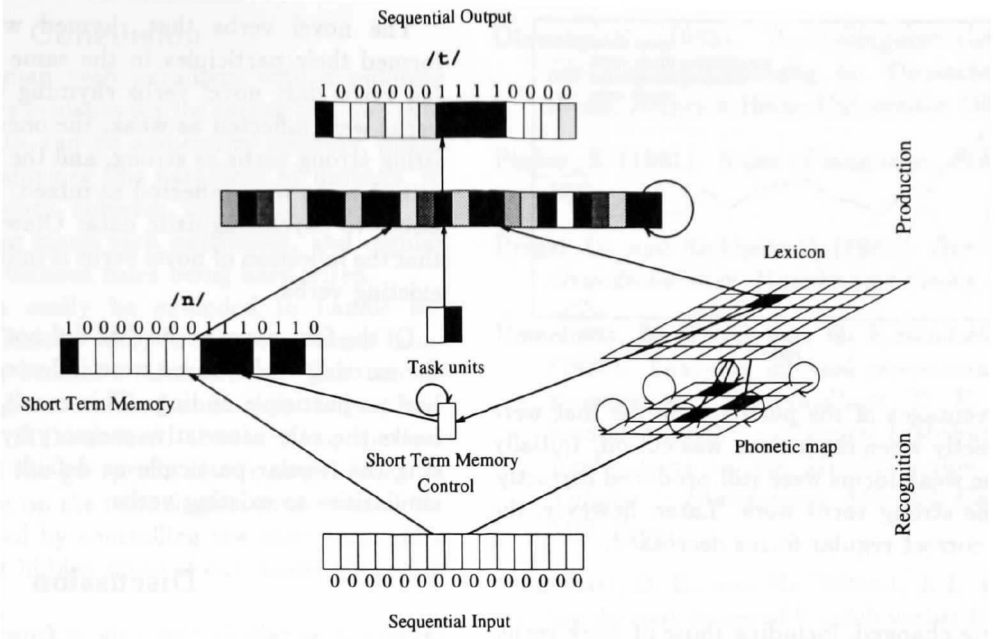

Figure 2: The network architecture.

In order to evaluate the role that the lexicon played in learning the task it was cut off and performance with and without the lexicon was compared.

Learning After training the network for 400 epochs it was able to produce all 100 infinitives and 93 of the participles correctly. All of the seven participles that were not learned were strong verbs having the phoneme /æ/ in the infinitive which was not changed correctly in the participle. This error might well be due to a local minimum on the error surface. Interestingly, the most frequent verb, **sein** with the participle **gewesen**, was among the unlearned.

The infinitive forms were learned well before the participle forms (after 40 epochs), and they were learned equally well with and without the lexicon. This was not surprising because the task of producing the infinitive simply consisted in duplicating the input sequence that was stored in the short-term memory.

The learning curve for the participle forms is shown in figure 3. Weak forms and strong participles that did not change their stems were learned equally well. Both mixed verbs and strong verbs that changed their stems were learned slower. Obviously it was easier for the network to prefix the participles with **ge-** where necessary and learn the correct participle ending than to change the stem.

In order to evaluate the role of the lexicon in learning the different participle forms, performance was also tested with the lexicon cut off. These results are shown in figure 4. Most significantly, at early training stages over 60% of the weak participles were produced correctly even without the lexicon (except that the prefix **ge-** was omitted because it relied on the lexicon to inhibit the short-term store, see above), but none of the strong forms were produced correctly, no matter whether they changed or kept their stems. This result strongly supports the dualistic framework which claims that regular (weak) participles are formed in the rule path but irregulars (strong) are produced with the lexicon. However, in contrast to the dualistic framework, the number of correct regulars decreased with further training. This indicated that the lexicon played a role even in the production of some weak forms at later stages of training. A closer analysis of this phenomenon showed that without the lexicon some parti-

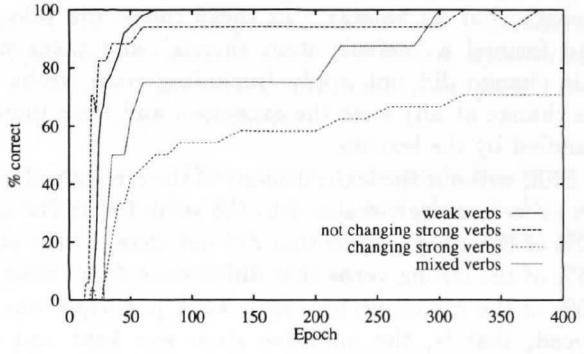

Figure 3: The learning curve for the participles of the different verb classes. The weak participles (52 types) were learned after about 60 epochs, the mixed participles (7 types) after 300 epochs, the strong participles that kept their stems (17 types) after 320 epochs, and the strong participles that required a stem change (24 types) were not fully learned.

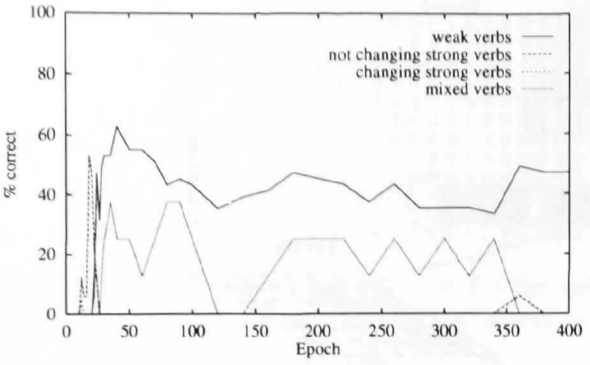

Figure 4: Percentages of the participle forms that were produced correctly when the lexicon was cut off. Initially over 60% of the weak forms were still produced correctly, but none of the strong verbs were. Later, however, the percentage of correct regular forms decreased.

ciple stems were changed, including those of weak verbs. For example, without the lexicon, the participle of the weak **hüpfen** (to jump) was not produced as **gehüpft** but as **(ge)hupft**, and the participle of **bewegen** not as **bewegt**, but as **bewagt**. In these cases, the rule path had formed a "default stem change" and cases where this change did not apply (including weak verbs with no change at all) were the exception and were therefore handled by the lexicon.

Still, without the lexicon many of the strong and mixed verbs were overgeneralized to the weak form: For about 80% of the strong verbs that did not change their stems, 25% of the strong verbs that did change their stems, and 30% of the mixed verbs now a weak participle was produced, that is, the infinitive stem was kept and a **-t** was suffixed. For example, the participle of **nehmen** (to take) which should be **genommen** was now **(ge)nehmt** and the participle of **sein** (to be) was **(ge)seint** instead of **gewesen**. Obviously the network had detected the correlation between not changing the stem and using the **-t** suffix. This correlation existed for the 52 weak verbs (67 tokens), while only 17 strong verbs (38 tokens) did not change their stems and used the **-en** suffix and only 7 mixed verbs (7 tokens) did change their stems and used the **-t** suffix.

The fact that the weak participle was overgeneralized to all other forms verified computationally that this form is the regular case in the German verb paradigm.

Generalization Generalization of the network model was tested with 16 non-existing novel verbs. Five of these rhymed with weak verbs, that is, they activated the same unit as a weak verb on the lexicon (e.g. **tolen** rhymed with **holen** and **pachen** with **wachen**). Three test verbs rhymed with existing mixed verbs, three with strong verbs that kept their stems, one with a strong verb that changed its stem, and four did not rhyme with any of the verbs.

The novel verbs that rhymed with existing verbs formed their participles in the same ways as the existing verbs did: novel verbs rhyming with existing weak verbs were inflected as weak, the ones rhyming with existing strong verbs as strong, and the ones rhyming with mixed verbs were inflected as mixed. This result corresponds to psycholinguistic data: Olawsky (1993) showed that the inflection of novel verbs is influenced by rhyming existing verbs.

Of the four novel verbs that did not rhyme with any of the existing verbs, three were inflected as weak and one had no participle ending. This result, too, qualitatively backs the rule–associative memory hypothesis in producing the regular participle as default when there are no similarities to existing verbs.

Discussion

The results of our experiments show that the modular connectionist network model was able to successfully learn the German verb paradigm, that it generalized to novel verbs in ways comparable to psycholinguistic data, and that its internal representations qualitatively corresponded to the rule–associative memory hypothesis (Pinker, 1991). Furthermore, the results gave computational evidence that the weak verb class is the regular class in German verb inflection. Our model deviates from the rule–associative memory hypothesis in two important aspects, however: First, the rule path and the associative memory are not strongly dissociated from one another, but both are needed to produce irregular and mixed verbs and even some of the regular verbs. The associative memory is needed to allow for stem changes and the **-en** ending of the irregular participles. Regularities for novel verbs that do not rhyme with existing verbs emerge because for these verbs the lexicon contributes nothing to the production of the participle, resulting in the short-term store keeping the infinitive stem and adding the suffix **-t** . The experiments further show that certain "micro-rules" are incorporated into the short-term path based on the statistical regularities of the input–output phoneme mappings, resulting in certain "default" vowel changes even for regular verbs when the lexicon is cut off.

The second difference to the rule–associative memory theory is that in our model no rules are hard wired but both the short-term store and the lexicon are entirely connectionist subsystems. The model's architecture is not specifically adapted to represent the characteristics of the German verb system, and its rule-like behavior emerges solely through the interaction of the short-term memory and the lexicon. This suggests that in humans, also, rule-like behavior can emerge without explicit rules being present.

Conclusion

Modeling the German verb paradigm with a modular neural network gave evidence for the theory that two subsystems, a rule-path and an associative memory, might be involved in handling of this paradigm by humans. In contrast to this theory, however, both paths worked together in producing many verb participles, and regular behavior emerged without rules being hard wired.

Our model can easily be extended to handle homophones: the lexicon only need to contain non-phonological, e.g., semantic information. This can be achieved by adding a semantic feature map to the phonological map (see Miikkulainen, 1990). Also, the fact that production of the **ge-** prefix in our model relies on the lexicon and not on the phonological properties of the verb can be avoided by controlling the short-term store from the recurrent hidden layer of our model instead of from the lexicon.

In principle, network models with similar architectures should be capable of handling all tasks that display default cases and exceptions, for example, script processing and algorithmic problem solving. The model, being able to manipulate symbols, might eventually lead to a unified theory of symbolic and sub-symbolic computing. This will be our main direction of future research.

References

Goebel, R. (1990). A connectionist approach to high-level cognitive modeling. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 852–859. Hillsdale, NJ: Erlbaum.

Goebel, R. (1991). Binding, episodic short-term memory, and selective attention, or why are PDP models poor at symbol manipulation? In Touretzky, D., Elman, J., and Hinton, G., editors, *Connectionist Models. Proceedings of the 1990 Summer School.* San Mateo: Morgan Kaufman.

Indefrey, P., and Goebel, R. (1993). The learning of weak noun declension in German: Children vs. artificial neural networks. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 575–580. Hillsdale, NJ: Erlbaum.

MacWhinney, B., and Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40:121–157.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., and Pinker, S. (1993). German Inflection: The Exception that Proves the Rule. Occasional Paper #47. MIT Center for Cognitive Science.

Miikkulainen, R. (1990). A distributed feature map model of the lexicon. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 447–454. Hillsdale, NJ: Erlbaum.

Olawsky, K. (1993). Psycholinguistische Experimente zur Partizipienbildung im Deutschen. Master's thesis, Heinrich-Heine-Universität Düsseldorf.

Pinker, S. (1991). Rules of language. *Science*, 253:530–535.

Pregel, D., and Rickheit, G. (1987). *Der Wortschatz im Grundschulalter.* Hildesheim: Georg Olms Verlag.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, 318–362. Cambridge, MA: MIT Press.

Rumelhart, D. E., and McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. In MacWhinney, B., editor, *Mechanisms of Language Acquisition.* Hillsdale, NJ: Erlbaum.

Ruoff, A. (1981). *Häufigkeitswörterbuch gesprochener Sprache: Gesondert nach Wortarten alphabetisch, rückläufig alphabetisch und nach Häufigkeit geordnet.* Tübingen: Niemeyer.

Wurzel, W. (1981). Phonologie: Segmentale Struktur. In Heidolph, K. E., Flämig, W., and Motsch, W., editors, *Grundzüge einer deutschen Grammatik*, 898–990. Berlin: Akademie-Verlag.