Explaining Language Universals In Connectionist Networks: The Acquisition of Morphological Rules

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 15(0)

Author
Lee, Chan-Do

Publication Date
1993

Peer reviewed

Explaining Language Universals In Connectionist Networks: The Acquisition of Morphological Rules

Chan-Do Lee
Department of Information Engineering
Taejon University
Taejon, 300-716, Korea
and
Center for Artificial Intelligence Research
Korea Advanced Institute of Science and Technology
Taejon, 305-701, Korea
email: cdlee@cosmos.kaist.ac.kr

Abstract

Across languages there are certain characteristics which they share. Linguists, trying to explain language universals, have come up with different theories: They argue for (1) the innateness of general linguistic principles, (2) the communicative functions reflected in linguistic structure, (3) the psychological demands placed upon language users, or (4) grammar-internal explanations. This paper tries to explain some of the morphological universals in the framework of a connectionist network, supporting the third approach. Employing simple recurrent networks, a series of experiments were done on various types of morphological rules. The results show that the model's performance mirrors the extent to which the different types of rules occur in natural languages. The paper explains how the model has discovered these universals.

1 Introduction

The study of language universals has been a major focus of modern linguistics for at least the past three decades. Why do languages share the universal properties that they do? Why do languages exhibit the range of variation that they do? Why are certain logically possible properties not found in any human languages? In attempting to answer these questions, it appears that linguists can be grouped according to four different theories, as explained by Hawkins (1988):

> Some argue for the innateness of general linguistic principles housed within a language acquisition device (LAD) which enables the new-born child to acquire the particular language of his/her community with remarkable speed and despite impoverished input. Others argue for a more social, rather than a biological, foundation to language: the communicative (discourse-pragmatic) functions that language users perform are reflected in linguistic structure. Yet others appeal to the psychological demands placed upon language users in the production and comprehension of language in real time. These so-called 'processing' demands are also argued to be reflected in its structure, as are certain intrinsic properties of our human perceptual and cognitive apparatus. Finally, there are more grammar-internal explanations, whereby one part of the grammar is claimed to be explained by another, for reasons essentially of internal consistency. (p. 3)

This paper tries to explain some of the morphological universals in the framework of a connectionist model. My approach in this paper is one that is based on the demands of learnability and processing (third approach from the above quote).

Employing simple recurrent networks, a series of experiments were done on various types of morphological rules. The results show that the model's performance mirrors the extent to which the different types of rules occur in natural languages. The paper explains how the model has discovered these universals.

2 Experiments

2.1 Method

I use a relatively constrained three-layer network, one in which feedforward connections are supple-

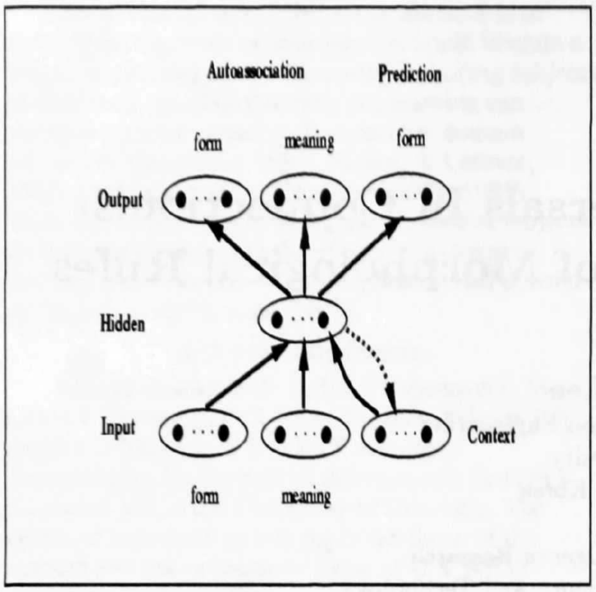

Figure 1: Architecture of the network

mented by limited feedback connections. Figure 1 shows the network architecture used for the experiments.

The architecture shown in Figure 1 is a slight modification of the simple recurrent network developed by Elman (Elman, 1990). Since morphological processes are temporal, we need to have some kind of short-term memory to store the previous events. The feedback connections from the hidden layer to the input layer serve this purpose.

The Form clique in the input and output layers consists of 8 units representing a phonological segment. Each Meaning clique consists of 7 units, 6 of which represent a stem meaning and 1 of which represents a grammatical feature of the input word (0 for the absence of that feature, e.g. singular, 1 for the presence of the feature, e.g. plural). The network has a variable number of hidden units and an equal number of Context units.[1] Each of the first two cliques receives input from the outside, while the Context units receive a copy of the activations on the hidden layer from the previous time step. The hidden units receive activations from the input layer, feeding the output layer. In addition, the activations on the hidden layer are fed back to the Context units. The output layer receives activations from the hidden layer. The output layer produces outputs in accordance with the current form and meaning and predicts the next input form. Given the current form and meaning, the network is trained to replicate them on one part of the output layer (autoassociation) and to predict what comes next in the sequence (prediction). My concern is with the arbitrary relationship between form and meaning; hence we need not concern ourselves in this paper with genuine semantics. The solid arrows denote the learnable one-to-many connections from the units on the lower levels to those on the higher levels. For example, any given unit on the input layer connects to all the on the hidden layer. The dashed arrow denotes the fixed one-to-one connections, on which no learning takes place with only one connection from a given higher level unit to a single lower level unit. There are no intra-level connections in the units in any clique or between cliques. The standard back-propagation learning rule (Rumelhart, Hinton, and Williams, 1986) is used to train the network.

This network has the capacity to associate form with meaning as well as form with form and meaning with meaning. Thus it can perform the task of the production of a sequence of segments given a meaning.

The results indicate that the network like this is capable of learning various types of morphological rules. That is, given training on the singular, but not the plural of *tone*, the network was later able to generate the appropriate plural suffix following the stem.

For example, the network was trained on pairs like the following:

```
(1)  ZONE  +  SINGULAR  -->  /zon/
(2)  ZONE  +  PLURAL    -->  /zonz/
(3)  TONE  +  SINGULAR  -->  /ton/
```

and then it was tested on pairs like the following to see if it then yielded correct phonological forms:

```
(4)  TONE + PLURAL  -->  /ton/ + ??,
```

where the items in capitals represent meanings.

Input words were composed of sequences of segments. Each segment consisted of a binary vector which represents modified Chomsky-Halle phonetic features (Chomsky and Halle, 1968): 1 for the presence of a particular feature and 0 for its absence, as shown in Figure 2. Each segment type was uniquely specified as a binary vector of 8 features.

There were 20 words for each simulation. Ten sets of randomly generated artificial words were used for each experiment. Twelve of these were designated "training" words, 8 "test" words. For each of these basic words, there was an associated inflected form. For convenience, I will refer to the uninflected

[1] The size of hidden layer was decided empirically, that is, pilot run was done for each experiment, and the network which gave the best performance was chosen.

```
features 8

vocalic high back anterior
coronal voice strident round

classification 16

p    0 0 0 1 0 0 0 0
b    0 0 0 1 0 1 0 0
t    0 0 0 1 1 0 0 0
d    0 0 0 1 1 1 0 0
k    0 1 1 0 0 0 0 0
g    0 1 1 0 0 1 0 0
f    0 0 0 1 0 0 1 0
v    0 0 0 1 0 1 1 0
s    0 0 0 1 1 0 1 0
z    0 0 0 1 1 1 1 0
i    1 1 0 0 0 0 0 0
e    1 0 0 0 0 0 0 0
a    1 0 1 0 0 0 0 0
u    1 1 1 0 0 0 0 1
o    1 0 1 0 0 0 0 1
#    0 0 0 0 0 0 0 0
```

Figure 2: Phonemes represented as binary vectors according to the feature matrix similar to that of Chomsky-Halle.

form as the "singular" and the inflected form as the "plural" of the word in question. The network was trained on both singular and plural forms of the training words and only on the singular forms of the test words. Words were presented one segment at a time. The network was trained on the autoassociation and prediction task. It was tested if it then yielded correct morphological forms.

To test the network's performance on the production task, I gave the network the appropriate segments for the stem successively, along with the meaning of that stem and the number unit on for plural. I then examined the prediction output units at the point where the plural morpheme should appear. Based on Euclidean distance, each output pattern was converted to the nearest phoneme.

2.2 Results

The morphological processes used were classified into three different categories: (1) addition, (2) deletion, and (3) mutation, and experiments were conducted on different types of rules; ones which are rarely found in human languages as well as ones which are commonly found, as shown below:

1. suffixation (+ assimilation): **fik -> fiks, gob -> gobz**
2. prefixation (+ assimilation): **fik -> sfik, gob -> zgob**
3. infixation (gemination): **ipa -> ippa**
4. initial deletion: **fik -> ik**
5. medial deletion: **ippa -> ipa**
6. final deletion: **fik -> fi**
7. reversal: **fik -> kif**

Table 1: Results of morphological process experiments. "% Segs Correct " refers to the percentage of all the segments which the network predicted correctly, while "% Affs Correct" refers to the percentage of correctly predicted affixes.

	% Segs Correct	% Affs Correct
Suffix	82.3	82.5
Prefix	62.0	76.3
Infix	73.5	42.5
Pre-del	12.5	
Mid-del	23.8	
Post-del	57.5	
Reversal	22.5	

The network succeeded on rule types which are common in human languages and failed on those which are rare or non-existent. Types 1 and 2 are common, types 3–6 less common, and type 7 non-existent. The results are summarized in Table 1.

3 Discovery of Universals

The model shows clear evidence of having learned morphological rules. The degree of mastery of the rules mirrors the extent to which the different types of rules occur in natural languages. In this section, I will explain how the model has discovered some universals.

3.1 Affixation and Deletion

The network performed much better on the affixation tasks than on the deletion tasks. The reason is that for the affixation tasks, the model had to predict the segment following the current phoneme, while for the deletion cases its task was to predict the phoneme that would come after the next one if it were not deleted. For the latter task, there is a gap between the current phoneme and the one that is to be predicted, making it more difficult for the network to predict the correct segment. Consider the following three cases:

```
(5)  /tap/ + /#/ and /tap/ + /s/
(6)  /tam/ + /#/ and /tam/ + /z/
(7)  /tap/ + /#/ and /ta/  + /#/
```

Problems (1) and (2) are somewhat easy, as in each case the identity of the last phoneme depends on the penultimate phoneme. The different contexts created by the penultimate phoneme are sufficient to ensure that different predictions can be made for the last phoneme. However, for (3), the final phoneme, that is, the word boundary, comes after the final phoneme in the stem for the singular case, while it comes after the second phoneme for the plural case. To predict the word boundary correctly in both cases, the network must develop different internal representations relative to the second phoneme for each case. (Indeed, the hidden unit activations have to be different if different outputs are to be produced.) Only in this way can the network generate the correct final phoneme (and in the next time step the word boundary) for the singular case and simply the word boundary for the plural case. Since the prediction tasks are the same for both cases up to the second phoneme, the network tends to develop the same hidden representations. This "homogenizing" process seems to strongly hinder learning in deletion tasks. Servan-Schreiber, Cleeremans, and McClelland (1988) report similar findings in their study on learning two arbitrary sequences of the same length and ending in two different letters such as:

PSSS P and TSSS T

They report:

> ...the predictions in each sequence are identical up to the last letter. As similar outputs are required on each time step, the weight adjustment procedure pushes the network into developing *identical* internal representations at each time step and for the two sequences – therefore going in the opposite direction than is required. (p. 29)

The very nature of the back-propagation learning rule and the structure of the model enable correct prediction of the affixed phonemes but make difficult the prediction of the segments after a phoneme is deleted. This is one explanation for the universal tendency of natural languages to exhibit many affixation processes, but few deletion processes.

3.2 Affixation and Assimilation

The network was able to generate the appropriate forms even in the prefix case when a "right-to-left" (anticipatory) rule was involved. That is, the fact that the network was trained only on prediction did not limit its performance to left-to-right (perseverative) rules since it had access to a static "meaning", permitting it to "look-ahead" to the relevant feature on the phoneme following the prefix. What makes this interesting is the fact that the meaning patterns bear no relation to the phonology of the stems. The connections between the stem meaning input units and the hidden layer units were being trained to encode the voicing feature even when, in the case of the test words, this was never required during training. For example, consider the following training set of artificial data.

```
(8)   FIK²  +  SINGULAR  -->  /fik/
(9)   FIK   +  PLURAL    -->  /sfik/
(10)  KOB   +  SINGULAR  -->  /kob/
```

When the network predicted the prefix of the word "KOB" for plural,

```
(11)  KOB + PLURAL  -->  ?? + /kob/,
```

it had available to it the characteristics of the first phoneme in the stem: among them notably the voicing feature. The meaning "KOB" has /k/ associated with it as its first segment. Thus the network knows that it has to produce /s/, since the grammatical feature unit is on and /k/ is voiceless.

In any case, it is clear that right-to-left assimilation in a network such as this is more difficult to acquire than left-to-right assimilation, all else being equal. Cross-linguistic studies of morphology have revealed an asymmetry in the frequency of affixing processes in favor of suffixing over prefixing (Hawkins, 1988), meaning that there are at least fewer opportunities for the right-to-left process.[3] I am unaware of any concrete evidence that would support left-to-right assimilation as easier than right-to-left assimilation, though in trying

[2] As explained before, the items in capitals represent meanings. Since the word is a made-up one, stem meaning is arbitrary.

[3] Of course, this does not necessarily mean that left-to-right rules are more common than right-to-left rules. In fact, right-to-left stress rules are more common than left-to-right ones.

to explain the asymmetry between the processes Hawkins and Cutler (1988) argue that:

> ...the linguistic and psycholinguistic evidence together suggest that language structure reflects the preference of language users to process stems before affixes, in that the component preferred for prior processing receives the most salient (initial) position in the word, the component to be processed second a less salient position. That is, the suffixing preference results in stems generally being ordered before affixes because language users prefer to process stems before affixes. (p. 311)

Why does the network find infixation most difficult among three suffixation cases? Consider the suffixation and prefixation cases. The affixes to be predicted is either /s/ or /z/, differing only in one phonetic feature, i.e. voicing. For the infixation cases, the task is to double the phoneme just seen, making the affixes different for each word tested. The variation of affixes to be predicted is much bigger in the infixation tasks than in the suffixation or prefixation tasks. Presented with the novel word, the network is prone to make more erroneous predictions in the infixation cases.

Whatever reason there might be, it is very encouraging to see that the model performs in a way that mirrors human language: suffixation is more frequent across languages than prefixation. and both are considerably more frequent than infixation.

3.3 Reversal

What is it that makes the reversal rule. apparently difficult for human language learners. so difficult for the network? Some aspects of the rule were learned. In 49% of the cases the network produced a CVC syllable as the plural form. What it could not do was to predict the correct consonants for the past tense.

Consider what happens in the suffix or prefix cases. The input consists of the sequence of phonemes representing the stem of a word, together with the stem meaning seen during training and the plural. not seen in this combination during training. Given a novel set of patterns, the network treats it as a combination of two sorts of patterns it has seen before: one of which is a sequence of phonemes representing the stem of a word, excluding the affix, along with the stem meaning; the other of which is the plural input, along with the feature of the segment that determines the appropriate plural form. The relevant phonetic feature is readily available in the suffix case as a part of the input. In the prefix case, as argued in the previous subsection, it is available as an acquired property of the stem meaning.

For the reversal case. if we think of the novel item in the form of a *set* rather than a sequence, then exactly the same set of segments is used for both singular and plural words. More importantly. however, since the network's task is to predict the next segment, there can be no sharing at all between the singular and plural forms in terms of prediction. Patterns on the hidden layer develop in response to prediction, so we should expect little similarity between context inputs for singular and plural words. As a result, the network does not have much material available for interpreting the novel reversed words. Presented with the novel plural form, it is more likely to respond based on similarity with a word containing a similar sequence of phonemes (e.g., `gip` and `gif`) than respond with the correct mirror-image sequence.

4 Limitations and Extensions

Despite its successes. this model is far from an adequate account of the explanation of language universals. Among many of the possible morphological processes, only a few that are typical word construction types and are also easy to implement were selected. Equally as important as the networks's ability to easily handle only those rules which occur in human languages. is the need for the model to explain more complex morphological phenomena. First, it should be able to account for *reduplication*. since the high frequency of reduplication in human languages implies that the model should be able to handle this kind of process. For the model to be able to produce words using this process. the current model may need to be substantially modified. Reduplication requires a primacy-oriented STM: such a model would have to focus on the beginnings of sequences because without the first segment. it would be very difficult, if not impossible, to predict the rest of the sequence of the word. Once a cue was input, the network would have to reproduce the part it had seen so far. A static representation of segments that encodes sequences, not just the contexts of the segments might be needed in this case. This static representation. then, could be used to reproduce the desired sequence when given the correct cue. The current model is designed to encode the contexts of segments, and its STM is therefore recency-oriented. since it is trained to predict the next segment; yet I am not sure if it would accommodate a static representation. The model might need some kind of help to acquire some knowledge of the syllable structure, as well as other features.

Another morphological process that is very challenging to the current model is that of *metathesis*, the transposition of two phonemes in a word. As demonstrated by the reversal experiment, this kind of process might be extremely difficult, if not impossible, for the model to acquire.

The experiments reported here were carried out on only a small, and severely restricted input corpus. Only 20 artificial words were considered in each simulation run. To be able to claim the plausibility of this model as an adequate system that can process morphological phenomena, I need to expand my research to a bigger data set. I used artificial language data for all the experiments, partly because it was simpler. One of the biggest drawbacks of this approach is that it cannot take into account the phonotactics of real human languages.

5 Conclusion

In this paper, some of the morphological universals were explained in the framework of a connectionist network. The study reported here is meant to model human learning and processing, attempting to understand language universals in terms of psychological demands. I do not believe that the work described in this paper necessarily makes strong claims that human perceptual processes are learned by the model used here (the model might not be the right one after all) and it remains to be seen how much my approach to the relatively trivial processes dealt with must be modified to deal with more complex processes and the elaborate mechanisms for handling them posited by traditional phonologists, but it gives an example of the kind of contribution connectionism can make to the search for language universals and their explanation.

References

Chomsky, N. and Halle, M. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Elman, J. 1990. Finding structure in time. *Cognitive Science* 14:179–211.

Hawkins, J. A. ed. 1988. *Explaining Language Universals*. Oxford: Basil Blackwell.

Hawkins, J. A. and Cutler, A. 1988. Psychological factors in morphological asymmetry. In J. A. Hawkins (Ed.), *Explaining Language Universals* 280–317. Oxford: Basil Blackwell.

Rumelhart, D., Hinton, G., and Williams, R. 1986. Learning internal representations by error propagation. In D. Rumelhart and J. McClelland (Eds.), *Parallel Distributed Processing, Volume 1*: 319–362. Cambridge: MIT Press.

Servan-Schreiber, D., Cleeremans, A., and McClelland, J. 1988. Encoding sequential structure in simple recurrent networks, Technical Report, CMU-CS-88-183, Dept. of Computer Science, Carnegie Mellon University.