Presented at the 1989 CUNY Sentence
Processing Conference (poster session)

– Clever Moms –

– Regularities in motherese that prove useful in parsing.

by

Cornell Juliano and Thomas Bever

University of Rochester

SUMMARY

All theories of language acquisition comprehension assume that initial hypotheses about sentential structure are available from lexical and sequential information. Learning and understanding a sentence both assume two major steps, isolating major phrases and assigning conceptual roles to the phrases. Our research addresses the extent to which segregation of phrases could be based on superficial patterns in linguistic input. In this discussion we first present the phrase segmenting results from a connectionist model which learns to segregate phrases from a superficial analysis of the input. Then we describe the empirical assumptions in the model and what they may reveal about language acquisition and comprehension.

RESULTS

We trained a connectionist model to recognize that the third of three words ends an utterance. The model was trained to converge on an output of '1' for the end of an utterance, and an output of '0' for non-terminal positions. After training on one set of texts, the model produced outputs between 0 and 1 for new texts, as Figure 2 shows.

The model's performance can be assessed in several ways. First, we can note that the mean activation level for actual utterance boundaries was higher than for non-terminal positions.

$$\text{utterance boundary average} = .477$$

$$\text{non-terminal position} = .119$$

Second, we can set boundary criteria using the average of all predictions in the generalized text. This value is equal to 0.2. This means that a boundary is considered as an activation level greater than 0.2. A nonboundary is considered as an activation level less than or equal to 0.2. With this criterion we obtain the prediction percentages in Table 1.

We can also examine the performance of the model on within utterance phrase-boundaries. This gives us a measure of the extent to which explicit training on ends of utterances generalizes to segregate phrases. The mean activation level for phrase boundaries is higher than that for non-boundary positions:

$$\text{within utterance phrase boundaries} = .301$$

$$\text{within utterance non-boundaries} = .100$$

If we again set the boundary criteria as the average of all predictions made in the "within utterance" generalized text we find the value of 0.125. We used this value to determine that a boundary is greater than 0.125 and a nonboundary is less than or equal to 0.125. This criterion produced the results seen in Table 2.

Finally, when we ran the generalization text looking at predictions for both "actual" utterance boundaries and "within" utterance boundaries

and used the 0.125 boundary criteria the model performed as shown in Table 3.

In brief, the model reveals a discrimination for phrase boundaries both in utterance final positions and within utterances.

The serial pattern of assignment of phrase-boundary activation levels affords an even more sensitive basis for segregating phrases. If we use the following two rules, the isolation of utterance-internal phrases is near-perfect:

1) a boundary occurs whenever the boundary activation level is lower than on between the previous position.

2) adjoin single-word phrases rightward, unless it is utterance-final, in which case adjoin it leftward.

These rules would operate on the sample output in Figure 2 to yield the phrases in Figure 5. A pictorial example of rule (1) is presented in Figure 1.

HOW IT WORKS

The important news is the simplicity of the information on which the model operates. The model examines up to three words of input at a time and is trained to predict whether or not the word in the right-most position ends the utterance. The overall schema of the model is given in Fig. 3.

The input to the model is presented in a serial fashion. The sequence of events is as follows. Information about the first word of an utterance is inserted into slot 3 and the simulator makes a prediction as to whether a phrase boundary follows the third slot. The information consists of the 3 things.

1. Is the word one the the 100 most frequent words that the model recognizes?

2. Word length, that is, short, medium, long, very long, greater than very long.

3. Distance in number of words between the current word and the last recognized word.

On the next presentation the word in slot 3 is shifted to slot 2 and the following word of the utterance is inserted into slot 3. A prediction is again made on whether a phrase boundary follows the third slot. This sequence is repeated, filling all three input slots and continues this shift/predict process until punctuation is encountered. The model then starts the sequence described above again.

A sample input and how it is presented to the model is demonstrated in Figure 5. Note that for each sequence in figure 5, the model makes a prediction that a boundary follows slot three.

The only positive training is when an utterance actually does end, as marked by punctuation. Insofar as the model fails to predict the end of an utterance, the activation weights are changed, according to standard back-propagation routines.

The model was trained on a 21600 word long sample of motherese from the Brown Corpus. The generalization text was 800 words of new motherese.

ASSUMPTIONS OF THE MODEL

The model assumes the ability to:

1.recognize the 100 most frequent words

2. recognize the length of all words

3. recognize the distance back to the nearest recognized word

4. recognize that an utterance has ended, and modify it's weights to predict an utterance boundary.

The success of the model in providing within-utterance phrase segmentation suggests that a superficial analysis of speech input can serve as an early phrase-segmentation stage in the acquisition and comprehension of language.

B = Boundary
NB = Nonboundary

	% Observed	
	B	NB
B	76	24
NB	18	82

Table 1 "Actual" boundary predictions

	% Observed	
	B	NB
B	71	29
NB	20	80

Table 2 "Within" boundary predictions.

% Observed

B = Boundary
NB = Nonboundary

	B	NB
B	82	18
NB	20	80

Table 3 "Actual" and "Within" utterance predictions.

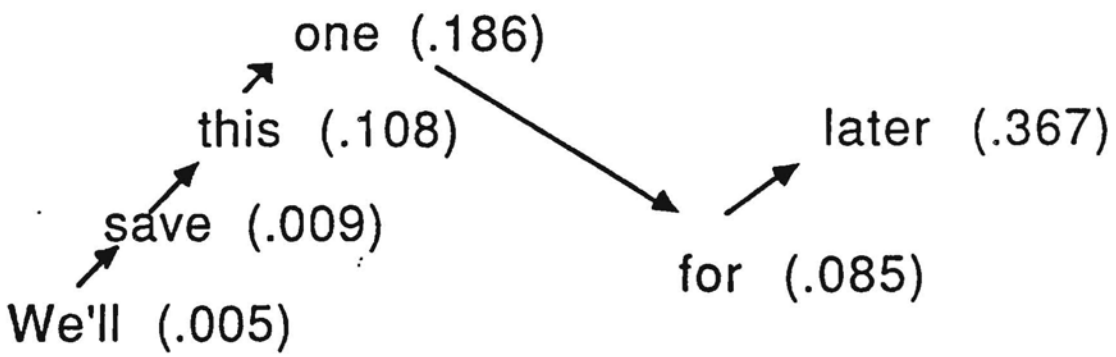
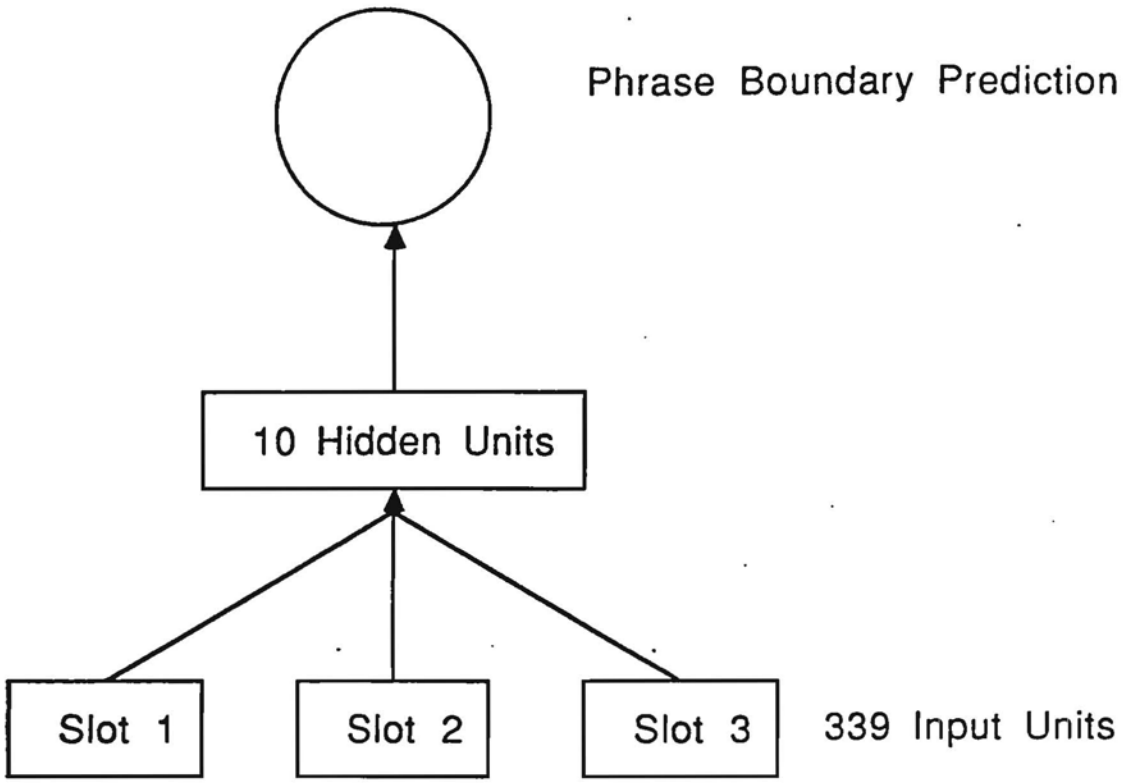


Figure 1 Application of Rule #1.



Each slot contains:
100 most frequent words (Recognized words, 100 units)
Word Length (5 units)
Distance to last recognized words (8 units)

Fig. 3

"Who (.036) is (.020) it (.498)?"

"We (.018) saw (.009) the (.073) children (.076) walking (.185)
in (.079) the (.055) rain (.755)."

"What (.257) did (.009) you (.057) hit (.012) that (.079) time (.333)."

"There's (.006) more (.131) in (.023) here (.501)."

"We'll (.005) save (.009) this (.108) one (.186) for (.085) later (.367).

"What (.257) are (.009) you (.073) cooking (.101) down (.087) there (.427).

"Children (.071) rain (.315), walk (095) rain (.366), yes (.875).

Figure 2. Sample of motherese with the activation level of predicted phrase boundaries included after each word.

[[Who] is it?]]

[[[We] saw the children walking] in the rain.]

[[[What] did you] hit that time.]

[[There's more] in here.]

[[We'll save this one] for later.]

[[[What] are you cooking] down there.]

[[[Children rain,] walk rain,] yes.]

Figure 4: Sample of motherese where phrases are segregated on the basis of rules 1 and 2.

Sequence	Slot1	Slot2	Slot3
1.	_____	_____	Adam_
2.	_____	_____	_put_
3.	_____	_put_	_the_
4.	_put_	_the_	truck
5.	_the_	truck	away_

Figure 5: A sample sentence, "Adam, put the truck away", is presented as input to the model in the sequential manner shown above. Note how the model restarts the sequence after the word "Adam". That is because "Adam" is followed by a comma, which is considered an end of utterance.