Assessing the Contribution of Representation to Results

Karen Anderson (KANDERS@CS.UCSD.EDU)
Jeanne Milostan (JMILOSTA@CS.UCSD.EDU)
Garrison W. Cottrell (GARY@CS.UCSD.EDU)
Computer Science and Engineering Department 0114
Institute for Neural Computation
University of California San Diego
La Jolla, CA 92093-0114

Abstract

In this paper, we make a methodological point concerning the contribution of the representation of the output of a neural network model when using the model to compare to human error performance. We replicate part of Dell, Juliano & Govindjee's work on modeling speech errors using recurrent networks (Dell et al., 1993). We find that 1) the error patterns reported by Dell et al. do not appear to remain when more networks are used; and 2) some components of the error patterns that are found can be accounted for by simply adding Gaussian noise to the output representation they used. We suggest that when modeling error behavior, the technique of adding noise to the output representation of a network should be used as a control to assess to what degree errors may be attributed to the underlying network.

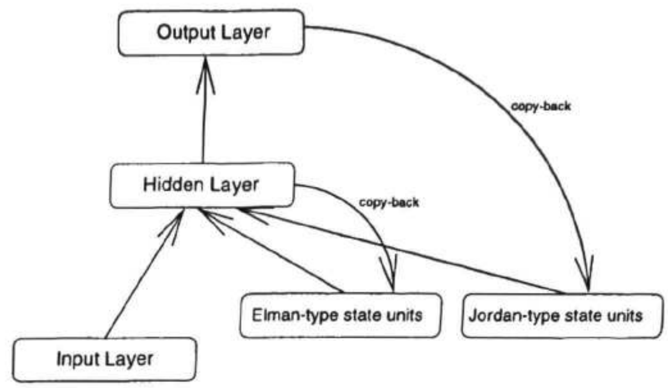

Figure 1: Pronunciation Network (Dell et al., 1993)

Introduction

Human error performance has often been cited as a window into the mechanisms underlying behaviors. Cognitive modelers, unlike artificial intelligence researchers, aim to have their models make the same mistakes people do, as well as account for correct behavior. They may then argue that, to the extent which the model matches both kinds of data, it is a better model than one that only accounts for correct performance. They may then be somewhat more confident in making the inference that, however the model works, humans may work the same way.

However, once a model makes error patterns similar to humans, it is important to understand what the source of those errors is in the model. In this paper, we replicate Dell et al.'s (1993) (henceforth Dell93) model of speech errors. While our error patterns are somewhat different than the ones they found, their error patterns are exhibited by some of our networks. On average, however, the performance of these networks do not match the human data as well as they did in Dell93. We attribute this to a large N for our models (we test fifteen networks of each type to Dell93's three).

This is not the point of this paper, however. Rather, it is a methodological one. Lachter & Bever (1988) criticized neural networks for using representations that predetermined the results. Of course, one *chooses* representations that are good for the domain. But how can we separate the importance of the representation in studies modeling error data? Here, we give a technique for assessing the contribution of the representation to the error patterns that separates it from the underlying network. Essentially, the technique is to add Gaussian random noise to the output patterns. To the extent that errors can be explained in this way, the underlying noise generating process (the network) is not very relevant. On the other hand, differences from this noise represent biases in the errors that can be attributed to the underlying network processing. In the following, we review Dell et al.'s model and results, we present our replications both with and without an underlying network, and discuss the implications.

The Model

Figure 1 shows the structure of the networks used in this work. This structure is identical to one of the networks used in Dell93 (they varied whether they had output recurrence or hidden recurrence or both). The network consists of a feed-forward path from input and state units through a hidden layer to a final output layer. Activation levels from both the hidden units and the output units are copied directly to their corresponding state units at the start of the subsequent time step.

The networks were trained to map from a representation of a word to pronunciation, using two types of input representation of the word. One consisted of a random pattern of bits, simulating the arbitrary nature of mapping from meaning to sound (Cottrell and Plunkett, 1991; Cottrell and Plunkett, 1995). The other input pattern simulated reading, by using three banks of five bits to represent the letters in the word, with a random code for each letter. In order to capture the temporal aspect of speech, the network output produces each phoneme segment sequentially, as in (Cottrell and Plunkett, 1991). The training set we used, identical to that from Dell93 (see Acknowledgments), consists of a subset of 50 frequent English words which have both 3 letters in their written form and 3 phonemes in their spoken form. Each word then is completed with a "null" end-of-word segment, so each output pattern is a sequence of four elements, always ending in the null pattern.

As in Dell93, networks were not trained to full competence; rather, training was halted at the end of any series of 50 epochs for which the errors on a per segment basis were less than 10%. This was done in order to analyze the types of errors made by the nets and to compare the frequency and types to those produced by humans in natural discourse settings.

In general, there are 3 types of errors which can occur: segment omission, insertion, and substitution. We classify these errors as to whether they violate any of the following constraints:

1. Phonotactic Regularity. Errors will usually produce sound sequences which are valid in the language (an estimated 1% of errors in humans violate this).
2. CV Category Effect. Vowels replace vowels; consonants replace consonants (violation rate 0.5%).
3. Syllabic Constituency. When a vowel *and* an adjacent consonant are in error, it is more likely to be a VC than a CV (estimated to be at a ratio of 3).
4. Initialness. Onset consonants are more likely to be in error than noninitial consonants (estimated at 62%).

Table 1 gives the rate at which these errors occur in human speech, as estimated by Dell93 from several error corpora and secondary sources (e.g., (Shattuk-Hufnagel, 1983; Shattuk-Hufnagel, 1987; Stemberger, 1983)).

Methods

We attempted to follow the methods outlined by Dell93 as closely as possible in training our networks. In the following, we give these procedures in detail as well as our methods for producing the noise model and scoring the errors.

The data consisted of the frequent word vocabulary from Appendix B of Dell93, which consisted of English words with both 3 letters and 3 phonological segments. Two input representations were used: a 30-unit *random* representation and a 15-unit *correlated* representation where each input letter was assigned an arbitrary 5-bit code. The networks were comprised of a feedforward path with 30 or 15 input units, depending on input representation, 20 hidden units, and 18 output units which were a feature-based representation of each phoneme. The output coding used is that from Dell93 Appendix A. The networks had as additional inputs a state layer which consisted of a copy of the output activations from the previous time step (hence it contained 18 units) and a context layer which consisted of a copy of the hidden layer activations from the previous time step (hence, 20 units). All connection weights were initialized to a random number in the range [-0.1,0.1], and each hidden and output unit had an additional bias input. The standard logistic activation function was used; momentum was 0.5 while the learning rate was 1.0 for the correlated input vector networks and 0.25 for the random input vector networks.

For each epoch of training, the words in the training set were presented in a new random order. At the beginning of each word, the context units were set to zero and the state units were initialized to the "null segment" (all values of 0.5). Each word's training input representation was clamped at the input layer while the output layer was trained to produce each phonemic segment of the word sequentially, followed by the null segment vector. Online backpropagation was used, with error propagated back and weights updated at the end of each forward pass of an output segment. An output segment was then considered correct if the values at the output layer were closer (by Euclidean distance) to the correct output than to any other valid segments.

15 separate networks of each of the two model types were trained, each starting with different random seeds. Performance of each network was examined at every 50 epochs; training was stopped when the networks achieved greater than 90% correct output segments. This occurred by 200 epochs for (Dell et al., 1993) and from 100 to 300 epochs for the replications presented here.

For the Noise simulations, a vector consisting of each output segment in the training set was perturbed with random noise generated by a Gaussian distribution with mean 0.0 and standard deviation of 0.31. This standard deviation was selected to produce overall segment correctness percentages consistent with those produced by (Dell et al., 1993) networks, approximately (but no less than) 90% correct segments. 15 noise sequences were produced, again each with a different random seed. Since the networks were constrained to perform at greater than 90% correct segments, any noise sequence which correspondingly caused more than 19 segment errors was eliminated from consideration and substituted with another sequence. Two such sequences were eliminated until the desired 15 sequences were produced. For both the noise model and the random- and correlated-input replication networks, each training set output segment was marked for type of error occurrence (omission, insertion, substitution). These errors were subsequently analyzed at the word level to determine how well they followed the word constraints. Phonotactic Regularity was assessed by hand-scoring the errors produced. Some ambiguous cases were given to Gary Dell for scoring (see Acknowledgments). Specific rules for error categorization (mutually exclusive but not exhaustive), labeled by the constraint they were used to measure, are as follows, scored in the following order:

1. *CV Category Effect:* A word has a cross category C-V error if the target output segment was a vowel and the produced output segment was a consonant, or vice versa.
2. *Syllabic Constituency Constraint:* A word has a syllabic constituent error if two immediately adjacent segments are in error, where one is a vowel replaced by another vowel and the other is a consonant replaced by another consonant. Further, any other consonant immediately adjacent to the vowel must not be replaced by an incorrect consonant. A syllabic constituent is then further categorized as a VC slip or a CV slip, depending on the relative location of the vowel in the pair.
3. *Initialness Constraint:* An error was categorized as a single-consonant substitution if the output segment produced a consonant different from the target segment consonant, and any immediately adjacent vowels were not replaced by an incorrect vowel. Each single-consonant substitution was then classified as initial or final depending on

whether the erroneous consonant preceded or followed the vowel within the same syllable. Finally, the consonant onset ratio is then the ratio of initial-consonant errors to the total of both initial and final consonant errors.

Since the purpose of this work is to attempt to model and explain human performance, we are interested in how well our models track human error data based on the constraints discussed earlier: Phonotactic Regularity, Cross-Category Errors, Consonant Initialness, and CV/VC Constituency. To this end, we use the deviation score described in Dell93, based on the differences in error proportions between the models' performance and human performance. Since we do not have variances, but only average data for humans, this measure tests for differences in means. For the overall deviation, the score is the sum of the squared differences between the human proportion and the model proportion. Each proportion is transformed by an arcsine conversion, included to correct "for differences in variability for extreme proportions" (Dell et al., 1993).

$$Deviation = \sum_i (arcsin(Hp_i{}^{0.5}) - arcsin(Mp_i{}^{0.5}))^2$$

where Hp_i and Mp_i are the human and model proportion of errors on constraint type i, respectively. i ranges over the five constraint types above. We are also interested in the deviation score for each constraint considered individually. The square roots of these scores were then compared using analysis of variance (ANOVA). Analysis of variance was also computed on the raw error proportions themselves.

Results and Discussion

Each pair of the 3 methods covered in this paper (random network input, correlated network input, and Gaussian noise-perturbed output units) was analyzed for difference in both overall deviation score and for deviation from human standards for each measure separately. The correlated and random networks performed significantly differently on the Phonotactic Regularity constraint, $F(1,28) = 9.134$, $p = 0.005$, with the correlated model deviating less from the human standard. These two models did not perform in a significantly different manner for the Cross-Category error constraint [$F(1,28) = 2.991$, $p = 0.095$], the CV Constituency [$F(1,28) = 0.178$, $p = 0.676$], the VC Constituency [$F(1,28) = 0.101$, $p = 0.753$], the Onset Consonant Ratio [$F(1,28) = 0.022$, $p = 0.884$], or the overall deviation measure [$F(1,28) = 0.119$, $p = 0.733$].

The Gaussian Noise model compared similarly with the random input model, with its smaller deviation from the human standard differing significantly for Phonotactic Regularity $F(1,28) = 19.631$, $p < 0.001$ while not differing in the other measures (Cross-Category $F(1,28) = 0.261$, $p = 0.118$, CV Constituency $F(1,28) = 0.060$, $p = 0.808$, VC Constituency $F(1,28) = 0.348$, $p = 0.560$, Onset Consonant Ratio $F(1,28) = 0.029$, $p = 0.865$, Overall $F(1,28) = 0.822$, $p = 0.372$).

The Gaussian-Noise model differed significantly from the correlated model on both Phonotactic Regularity [$F(1,28) = 4.816$, $p = 0.037$] and on Cross-Category errors [$F(1,28) = 17.808$, $p < 0.001$] with a closer match to the human standard for both measures; it was not different on the CV Constituency [$F(1,28) = 0.326$, $p = 0.572$], the VC Constituency [$F(1,28) = 0.068$, $p = 0.796$], the Onset Ratio [$F(1,28) = 0.001$, $p = 0.979$], or the overall deviation measure [$F(1,28) = 0.371$, $p = 0.547$].

Using the deviation score measure, it would appear that all three techniques performed approximately the same on most of the measures. However, examining the raw average error proportions gives a different picture, as shown in Table 1.

One important measure to consider in Table 1 is that of the Onset Ratio. Based on the deviation score measure used in Dell93, the error proportions for each of the 3 simulation methods presented here appear nearly equivalent. However, inspection of the raw scores makes it clear that this is not the case: the random and noise model are much less than, and the correlated model much greater than, the human standard of 62%. This distinction does not present itself in the ANOVA results because the measure used calculates the *squared deviation* from the standard. This type of measure is intended to ensure that data with high variance centered around a mean receives a worse score than data centered on that mean which has small variance. The systematic offset to one side or the other can then potentially be lost in this type of measure, as can be seen above. Clearly, use of the deviation measure was necessary in the original case, since the only data available from the human standard is the means, and the data sets could not be directly compared in the usual manner without some form of variance measure on the human standard. However, in this case we actually do want to compare the three models with each other so we can dispense with the deviation score and compare the raw error proportions with analysis of variance.

For this measure, the analysis appears a bit different. For the Phonotactic Regularity constraint the random model differs from both the noise ($F(1,28) = 18.445$, $p < 0.001$) and the correlated model ($F(1,28) = 7.439$, $p = 0.011$), and the noise and correlated models also differ ($F(1,28) = 6.886$, $p = 0.014$). For the Cross-Category error constraint, the noise and correlated models produce different results ($F(1,28) = 17812$; $p < 0.001$). Neither the noise and random ($F(1,28) = 2.574$, $p = 0.120$) nor the correlated and random ($F(1,28) = 2.836$, $p = 0.103$) differ. For the CV and VC Constituent measures, no significant differences were found: noise and random ($F(1,28) = 0.154$ for CV & 0.000 for VC, $p = 0.698$ & 0.997); noise and correlated ($F(1,28) = 1.799$ & 0.157, $p = 0.191$ & 0.695); correlated and random ($F(1,28) = 3.035$ & 0.136, $p = 0.092$ & 0.715). In the case of the Onset Ratio, the correlated model differs from both the random model ($F(1,28) = 36.390$, $p < 0.001$) and the noise model ($F(1,28) = 50.835$, $p < 0.001$). The noise and random models do not differ, ($F(1,28) = 0.455$, $p = 0.506$).

What does this maze of statistics tell us? Recall that what we want to know is how much the representations chosen affect the outcome of the simulations. Thus, we are concerned with how the noise model performs in terms of the human performance standard, and then how the network models add to that.

Table 1: Model Error Characteristics

Model	Phonotactic Regularity	Cross Category	Onset Ratio	VC Slips	CV Slips	VC/CV Ratio
Random	86.0	97.8	37.9	4.93	1.73	2.85
Correlated	95.5	95.3	83.4	4.25	4.00	1.06
Noise	99.7	99.7	32.2	4.94	2.19	2.26
Dell93						
Random	98.0	100.0	58	10	4	2.5
Correlated	94.3	100.0	62	9	2	4.5
Human	99.0	99.5	62	6	2	3

Analysis: Phonotactic Regularity

The noise model produces only one phonotactic regularity violation, thus averaging 99.7% across the 15 noise runs. This high level of correctness corresponds well with the human error data estimated at 99%, allowing attribution of performance on this measure to the representation chosen.

Analysis: Cross Category Constraint

As alluded to in Dell93, the high conformance to the Cross-Category constraint is mostly due to the output representation. Cluster analysis of the output representations show that the vowel representations are distinct from the consonant representations, and thus small deviations are more likely to change a vowel to another vowel or a consonant to another consonant. This is empirically verified by the performance of the noise model: vowels do not, in general, become closer to consonants with small errors and vice versa. Thus, the Cross-Category constraint is attributable to the representation rather than the network structure or learning. In fact, the network models move the errors further from the human data.

Analysis: Onset Ratio

Analysis of the Onset Ratio constraint is more complex. Recall that this constraint claims that segment onset consonants are more likely to be in error than segment coda consonants. Both human standards and some of the networks in Dell93 show this. However, the noise model examined here shows the opposite effect, with coda consonants being in error almost twice as much as onsets. This phenomena can be explained by confluence of two statistical factors. First, the output phoneme data set is biased toward coda consonants. There are 40 consonant phonemes in the onset category, compared to 54 in the coda set. This produces an onset to total consonant ratio of 42%. If each consonant was equally likely to be in error, this is what the Onset ratio should be. One would then expect the noise model to have produced this ratio, but it did not. For the total of 15 instances of the noise generation, there were 34 consonant-to-consonant errors in the onset position with 66 in the coda position (a ratio of approximately 34%). However, there are several consonants which occur in the coding of the output phonemes, which do not occur at all in the training set. Of specific interest are the 4 syllabic consonants, commonly depicted as /rS/, /lS/, /nS/ and /mS/. These syllabic consonants have representation identical to their "plain" consonant partners except in the case of the voicing feature. That is, /rS/ differs from /r/ in having one additional bit turned on. All of the syllabic consonants have this same difference, and additionally no other consonants include this feature. This means that there is a bit turned on in the coding of the syllabic consonants which is not used in any other consonants.

Examination of the error set for the noise model revealed 21 cases where a target consonant produced a syllabic consonant in error; in each of these cases, the intended consonant was the corresponding "plain" version of the consonant (In all 30 networks of the random and correlated models, there was only one error of this type total). Since an output segment is classified as the phoneme to which it is closest in Euclidean distance, none of these phonemes would have been marked in error if the syllabic consonants were excluded from the set. The majority of the syllabic errors which did occur were targets of /r/ and /n/, both of which occur more frequently in the coda position. Eliminating these from the total error counts mentioned in the above paragraph, we then are left with 24 onset consonants and 56 coda consonants in error, for an Onset Ratio of 44%, much more in line with the predicted 42%.

This being accounted for by the representation, it can then be concluded that the difference between the above rates and the performance of the correlated model (and the analogous networks from Dell93) on the Onset Ratio constraint is due to the sequential nature of the networks.

Analysis:Syllabic Constituency

The Syllabic Constituency Constraint (VC/CV Ratio) derives from the observance of predominantly more VC than CV type substitutions in human error data. The respective frequencies cited by Dell93 are 6% and 2% of all segment errors. While analysis of the deviation scores and raw error proportions revealed no significant differences in either the VC or CV substitution error rates between any pair of the three model types, it also did not show any of them to be different from the human data. On the whole, given that all of the models were constructed to make errors, it is not surprising that some percentage of them should make errors on adjacent segments, nor that some of these errors should be VC or CV substitutions. As previously noted, the segment representation is such that the easiest errors to make are $V \rightarrow V$ and $C \rightarrow C$ substitutions. The interesting phenomena is, rather, that the predominance of VC over CV substitutions observed in the human data is maintained in the models' errors. The VC to CV ratio for the human data is 3:1; those for the correlated input, random input, and Gaussian Noise models are, respectively, 4.9:1.7, 4.3:4.0, and 3.1:1.9. Though none of the models' ratios match the human ratio precisely, the random input and Gaussian Noise models show a strong preference for VC errors. The correlated-input model's VC slips barely outnumber its CV slips, however.

If not for the fact that the Gaussian Noise model also exhibits this bias toward VC slips, this evidence would seem to suggests that at least the random input network model may have captured some facet of the human production process – for how can the representation of the individual segments possibly explain the preference for erring on adjacent segments solely because the vowel precedes the consonant? In discussion Dell93 attributed the presence of the syllabic constituency effect in their networks' errors to an interaction between the training vocabulary and the sequential nature of the networks. Pointing out that the training vocabulary has a greater redundancy in its VC than in its CV patterns, Dell93 conjectured that this caused the networks to develop such a strong association between the segments of the oft repeated VC sequences that they came to represent them internally as a single unit. Hence, when an error occurred by chance on the initial vowel of an infrequent VC pair and resulted in another vowel that was part of a frequent VC pair, the network was apt to be pulled off its original course and onto the well-worn trajectory between the in-error segment and its compatriot consonant. This seems like a sound enough explanation, but we conjecture that it is only part of the story.

Since the Gaussian Noise model has a strong tendency to produce VC over CV errors, even more so than one of the neural network models, we are forced to look harder for the source of this predisposition. Looking at the training vocabulary from a slightly different angle than Dell93, a second potential explanation for VC to CV predominance becomes apparent.

In Table 3 below it can be seen that there are simply more opportunities for a VC error to be made than for a CV error. To assess the contribution of this imbalance toward inducing a greater number of VC errors, we calculated the expected number of CV and VC errors based on the token type probabilities and a random distribution of segment errors. The probability of a segment error occurring on any given segment was based on the average number of segment errors for all models combined. The expected error rates obtained were 0.00475 and 0.00434 for VC and CV errors, respectively. Using these predictions we then estimated the expected VC and CV error proportions as 0.0684 and 0.0629.

The expected proportions above, however, show only a slight bias toward VC errors. What then can be responsible for the Gaussian Noise model producing nearly double the amount of VC errors as CV? Since the effect cannot be entirely explained by token type frequency with the assumption that errors are randomly distributed, perhaps this assumption is fallacious. The only reasonable explanation remaining is that there is a consistent and qualitative difference between vowel-preceding and vowel-succeeding consonants. For some reason, vowel-succeeding consonants must be predisposed at the representational level to errors.

To investigate this possibility, we analyzed the consonants that follow a vowel compared to the ones that precede a vowel. In particular, we examined the number of neighbors each kind of consonant has within a Hamming distance of 3. As seen in Table 2, the consonants following a vowel have many more neighbors, on average, than the consonants preceding a vowel. This means that consonants following a vowel are much more likely to slip to an erroneous consonant, which, if their neighboring vowel also slips, would lead to more VC errors than CV errors.

Table 2: Frequency & Neighbor Density of Vowel-adjacent Consonants

Pre-vowel	Freq	# Neighs	Post-vowel	Freq	# Neighs
h	5	2	t	14	8
g	5	2	n	12	8
r	4	8	r	6	8
b	4	8	d	6	11
y	3	2	z	4	10
w	3	2	s	2	8
s	3	8	g	2	2
n	3	8	m	1	7
m	3	7	l	1	8
k	3	2	b	1	8
l	2	8			
f	1	6			
t	1	8			
p	1	7			
j	1	5			
d	1	11			

Again, Dell93 claim that the greater redundancy of VC versus CV units in the frequent vocabulary is responsible for creating the effect of the syllabic constituency constraint in their models. Presumably, the VC units come to be considered a single unit and the VC slip predominance is a resultant "emergent property". Our explanation above suggests that the Gaussian Noise models that we produced can account for this effect. However, there is a noticeable qualitative difference between the VC slips of the network models and the VC slips of the Gaussian Noise model. Dell93 cited the following three errors as particular examples of VC errors that indicated the networks were learning to associate redundant VC pairs :

big ⟶ /bed/
him ⟶ /h3n/
old ⟶ /#nd/ (#=schwa)

In all the cases above, the low frequency VCs : /Ig/, /Im/ and /ld/ are replaced with VCs that occur more that once in the training set. These very same errors also showed up in our network models' output – some, exactly as listed above, and others with just the relevant VC → VC substitution replicated. Furthermore, looking at the frequencies of replaced vs. replacing VC units for all such errors made by our models shows that the random model produces VC errors that are frequent in the target vocabulary, while the correlated model and the Gaussian model do not. In particular, the Gaussian

Table 3: Vocabulary syllable structure

Target V-C Pattern	CCC	CCV	CVC	CVV	VCC	VCV	VVC	VVV
Frequency in Vocabulary	0	1	41	0	6	2	0	0

model takes *frequent* pairs and replaces them with pairs that *never* occur.

The conclusion we draw from this is that, while the network models have the VC/CV ratios in the right ballpark, they cannot claim to have them because of an effect of the underlying network. Rather, all that can be claimed is that the content (which is much more human-like), not the proportion, of the errors is due to the network's action.

Conclusion

While this work is still at a preliminary stage, we believe we have demonstrated that an important component of any analysis of network error must separate out the contribution of the network from the representation. In the above analysis we found, for example, that several of the components of the error pattern could simply be accounted for by adding noise to the output representation. In this case, the network's role is simply as the supplier of that noise.

Acknowledgments

We thank Cornell Juliano for his computer archeology research that led to providing us with the input/output patterns from the 1993 network models. We thank Gary Dell for providing some of the error data from the same model, answering our seemingly endless questions concerning the error scoring procedures, and for actually scoring some of the errors for us! Thanks also to Gary's Unbelievable Research Unit (GURU) for helpful comments on this work.

References

Cottrell, G. W. and Plunkett, K. (1991). Learning the past tense in a recurrent network: Acquiring the mapping from meaning to sounds. In *The Thirteenth Annual Conference of the Cognitive Science Society*, pages 328–333, Chicago, IL. Lawrence Erlbaum, Hillsdale.

Cottrell, G. W. and Plunkett, K. (1995). Acquiring the mapping from meaning to sounds. *Connection Science*, 6(4):379–412.

Dell, G. S., Juliano, C., and Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, 17(2):149–196.

Lachter, J. and Bever, T. G. (1988). The relation between linguistic structure and associative theories of language learning: A constructive critique of some connectionist learning models. *Cognition*, 28:195–247.

Shattuk-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In MacNeilage, P., editor, *The Production of Speech*. Springer-Verlag, New York.

Shattuk-Hufnagel, S. (1987). The role of word-inset consonants in speech production planning: New evidence from speech error patterns. In Keller, E. and Gopnik, M., editors, *Motor and Sensory Processes of Language*. Erlbaum, Hillsdale, NJ.

Stemberger, J. (1983). *Speech Errors and Theoretical Phonology: A Review*. Indiana University Linguistics Club, Bloomington.