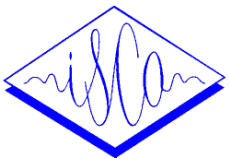

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PHONOLOGY AS A BYPRODUCT OF LEARNING TO RECOGNIZE AND PRODUCE WORDS: A CONNECTIONIST MODEL

Michael Gasser

Indiana University

ABSTRACT

This paper investigates the possibility that phonological knowledge emerges out of learning how to process words. A connectionist model of word recognition and production is presented, and a series of experiments is described in which a network is trained to recognize or produce a small set of words from an artificial or a real natural language. In the process of learning these tasks, the network develops internal, distributed representations of its state at different points during processing. In one set of experiments, the internal representations which emerge during a recognition task are treated as inputs to other networks, where their adequacy as syllable representations is tested. It is shown that the representations (1) support word production as well as recognition, (2) support mutation, insertion, and deletion processes, and (3) are robust to noise in the input. In another experiment, representations which develop during a production task are analyzed using a dimensionality reduction technique. It is shown that two of the dimensions exhibit some of the properties of the tiers of autosegmental analyses.

10.21437/ICSLP.1992-88

WORD RECOGNITION, WORD PRODUCTION, AND PHONOLOGICAL REPRESENTATIONS

Consider a simplified version of the task a child faces in learning to recognize and produce words, one in which the problem is to learn the mapping between sequences of phonetic segments and sets of lexical entries and grammatical morphemes. There are several ways in which the child's task might be solved. One possibility, naive from a linguistic perspective, is to map each sequence directly onto the word which it represents. But this is clearly inadequate given the productivity of morphological processing: how is the child to recognize a sequence representing a combination of morphemes which s/he has never heard before or to produce such a novel sequence? One is led to believe in representations which mediate the segment-sequence-to-morpheme mapping, phonological representations in which morphemes have an invariant form. Further, in order that generalizations can be made from recognition to production, these mediating representations should be shared by the two processes.

Even in a language with no complex morphology, where we don't expect productivity, there are arguments for intermediate phonological representations. Word recognition and production require a short-term contextual memory, yet it is inefficient to store the raw signal itself. Rather the cognitive system apparently makes use of the regularities in the input to build more compact, abstract context representations. Attention to regularity also allows the recognition system to correct for errors or noise and to focus on what is difficult by ignoring what is predictable [4]. For production, knowledge of regularity reduces memory load by permitting compact lexical representations and enables hierarchical planning [11].

There is nothing very controversial in suggesting that there is a level of phonological representation in the language processing system. The mediating representations I have been discussing are in fact similar to phonologists' underlying representations. However, they differ in that they are meant to actually arise during processing. That is, they represent part of the output of a learning/processing model rather than the input to a phonological derivation.

In sum, word recognition and word production require one or more levels of intermediate phonological representations which are shared by the two processes. The question addressed in this paper is how such a system could evolve as it is presented with the supervised task of learning to recognize and produce words. I will describe a connectionist architecture designed to perform the recognition and production tasks which develops phonological representations on its hidden layer. Specifically, in one set of experiments, a network develops syllable representations which support production as well as recognition, which are robust to errorful input, and which support systematic phonological transformations. In a second experiment, a network's hidden-layer representations are shown to be organized around dimensions corresponding roughly to the tiers of an autosegmental analysis.

SEQUENTIAL NETWORKS

A connectionist **pattern associator** [13] is a feedforward network trained to map inputs onto outputs. Usually one is interested in whether a pattern associator can make generalizations, that is, whether it can respond with appropriate outputs given inputs it has never seen before. The task at hand involves the association of patterns, and we would like to know whether the processing system is capable of recognizing or producing novel sequences. However, the sequential nature of the task makes a simple feedforward network inadequate because such a network does not afford the temporal context that is required.

A feedforward network gains a context, and the capacity to process sequences, however, with the addition of recurrent connections on its hidden and/or output layers [3, 10]. Figures 1 and 2 show two such sequential network architectures which have proven successful for the word recognition and production tasks as defined above. In the figures boxes represent layers of connectionist processing units and solid arrows complete connectivity between layers. All of the connections are modified during training using the back-propagation learning algorithm [15].

The recognition network is presented with a segment on each time step (and a boundary segment at the end of a word) and trained to output the pattern on the LEXICON and GRAMMATICAL MORPHEME layers which corresponds to the input word. There is a single unit for each morpheme on these layers. The network is also trained to auto-associate the input segment. This forces it to distinguish the segments early in training. The pro-

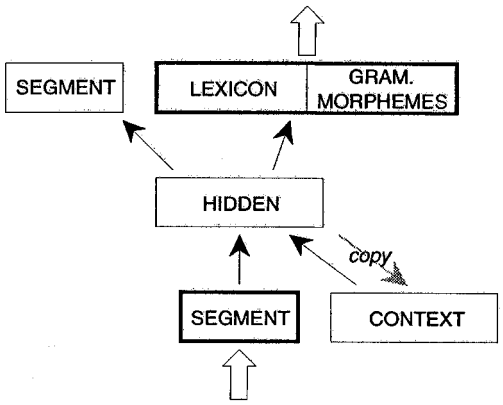

Figure 1: Sequential Network for Word Recognition

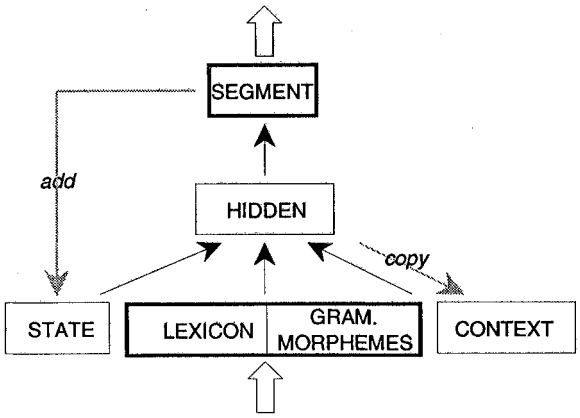

Figure 2: Sequential Network for Word Production

duction network is given a constant LEXICON and GRAMMATICAL MORPHEME input and trained to output in sequence the segments for the appropriate word.

Both networks include a layer of CONTEXT input units, which are activated with the pattern which appeared on the hidden layer on the previous time step. Thus the hidden layer has recurrent connections. Because these connections are trainable, the system can learn to develop context representations to solve the task it is given. In addition to the CONTEXT layer, the production network has access to its previous outputs via a STATE layer. On each time step the pattern on the output SEGMENT layer is added to a decayed version of the pattern on the STATE layer. For production, this combination of two kinds of recurrence has proven to be superior to either kind alone.

LEARNING SYLLABLE REPRESENTATIONS

Word Recognition and Syllable Representations

We have seen that it is desirable for phonological representations to be usable in both recognition and production. One way to ensure this is to use the hidden-layer patterns from a trained recognition network as inputs to a production network. If these hidden-layer patterns constitute **summary** representations of the sequences that they follow, then a production network might be trained to output these same sequences given the patterns as inputs. In the experiments that follow, hidden-layer patterns were saved following the presentation of each input syllable. It was hoped that these patterns would constitute intermediate phonological representations which embodied the structure of the syllables, albeit in a distributed form [16].

Stimuli for these experiments consisted of sequences of phonetic segments in an artificial language. Segments were represented by vectors of 11 phonetic features. Syllables in the language were characterized as follows: ONSET $\rightarrow$ $\{0,p,f,m,t,s,n,k,x\}$, NUCLEUS $\rightarrow$ $\{i,e,a,o,u\}$, CODA $\rightarrow$ $\{0,n,s\}$. Each experiment began with the training of a recognition network to categorize a set of words in the language. Each word consisted of two legal syllables, and the set of training words was generated by randomly combining pairs of syllables, with the restriction that no identical pairs were included. Note that there was no GRAMMATICAL MORPHEME layer for these experiments. Once the recognition network had been trained, representations for each of the 135 possible syllables, consisting of hidden layer patterns following the presentation of the syllable sequences, were extracted from the network. These syllable representations were then used as inputs to other networks.

Syllable Representations for Word Production

The first experiment tested whether the syllable representations were usable for word production. First 100 two-syllable words were generated, resulting in a set which contained 104 of the 135 possible syllables in the language. Next the recognition network was trained to identify the segment sequences representing the words. Performance on word recognition following extensive training was relatively poor: only 17 of the 100 words were correctly identified. Still it was felt that in attempting to learn to distinguish the words, the network might have developed distinct representations for the syllable sequences that made them up. Representations for all 135 possible syllables (including those not trained on) were set aside by presenting the network with the syllable sequences and then saving the final pattern on the hidden layer.

Next these syllable representations were used as inputs to a production network like that shown in Figure 2. The syllable inputs replaced the LEXICON and GRAMMATICAL MORPHEME layers. 20% of the syllables were randomly selected to be set aside for testing the network for generalization. The production network was trained to output each training syllable sequence followed by a boundary symbol. Following training, the network made errors on only 7 of the 95 segments in the test sequences (7.3%; chance: 85.3%).

These results indicate that the recognition network is able to generalize about syllable structure on words containing a subset of the possible syllables and that the distributed representations developed during training can be used for production.

Robustness of the Representations

A further question to be asked about the syllable representations is whether they are robust to noise and errors. To test this, the trained recognition network was presented a representative set of 142 bogus syllables, sequences which did not conform to the language the network had been trained on. These included sequences with segments not among the phoneme inventory (e.g., *b*), illegal codas (e.g., *fap*), long nuclei (e.g., *mua*), cluster onsets, and no nuclei. The hidden-layer representations for these sequences were saved and presented to the trained production network. The output of the production network was then examined to determine whether the networks would in effect correct the inputs. The production network responded to 97 of the 142 sequences (68%) by replacing the original sequence with a legal syllable in the language. Typical responses included the follow-

ing: *kn* → *ken*, *kfe* → *ke*, *xou* → *xu*, *pik* → *pi*, *zan* → *nan*.

These results are further evidence that the networks have learned to represent the structure of syllables. Because input segments such as /b/ are replaced by legal ones such as /p/, they also indicate that the networks have learned about the phoneme inventory of the language.

Transformations on the Representations

Finally, if the syllable representations really encode phonological structure, they should support phonological operations. To test for this, a set of feedforward networks was trained to take the syllable representations from the recognition network and output the syllable representation that would result when applying a particular rule to the input syllable. Four rules (and four networks) were used: one which replaced the syllable nucleus with *u*, one which replaced the coda with *-s*, one which replaced the onset with the fricative in the same place of articulation as the onset of the original syllable (or by *s* if there was no onset), and one which deleted the coda (if there was one). Each network was trained on 80% of the syllables, then tested on the remaining 20%. For each rule, over 95% of the test syllables were generated correctly (chance: 0.7%).

These results indicate that the representations learned by the recognition network encode syllable structure in a way which makes it accessible to mutation, insertion, and deletion.

LEARNING TIER-LIKE REPRESENTATIONS

One of the major recent advances in linguistics has been the view that phonological and morphological units exist on separate tiers [5]. Each tier contains autosegments which during derivation are associated with positions on a skeletal tier, where the final output sequence is specified. Phonology and morphology are now multi-dimensional, and it is reasonable to ask whether tier-like effects can be observed in the multi-dimensional distributed representations that emerge in the networks I have been describing.

Semitic languages, which were an early inspiration for autosegmental analyses [10], provide an especially clear illustration of what we are looking for. In these languages, verb roots consist of sequences of consonants whose position in surface verb forms depends on the structure of a phonological template associated with some grammatical category. For example, in the Ethiopian Semitic language Amharic, three of the forms based on the root *sbr* 'to break' are *tisbʌr* 'let her break' (jussive), *sʌbbʌrʌčč* 'she broke' (perfect), and *sʌbra* '(she) having broken' (converb). Each of the three aspects which distinguishes these words defines a template within which the root consonants are placed, and person-number-gender affixes are added to the resulting stem. In an autosegmental approach, the root consonants, aspect template, and person-number-gender affixes appear on separate tiers, and during derivation the autosegments on these tiers are associated with positions on the skeletal tier.

How might such an analysis translate into the behavior of a sequential network? Since a recognition network cannot know which template it is seeing early in the presentation of a word, I will be concerned here only with production. As the segments in a verb form are output during production, the system's representation of its evolving state is just the pattern on the hidden layer. If the system is moving through several independent tiers simultaneously, we would expect this to be reflected in the dimensions observable in the hidden layer. Thus for Semitic verbs, there should be a dimension with characteristic regions or directions of movement associated with the output of the root consonants, independent of what these consonants are and where they appear in the final sequence.

In the simplest case, these tier-like dimensions would be represented by individual hidden-layer units. A further possibility, pursued here, is that by rotating the axes of hidden-layer space, we can find a relatively small set of dimensions which better characterize the hidden layer patterns and which include the dimension we are looking for. **Principal component analysis** is a technique which for a given set of data vectors yields a set of orthogonal vectors, or components, which are ranked in terms of how much of the variance in the data they account for. This technique has proven useful in analyzing the behavior of sequential networks [3, 12].

Learning to Produce Amharic Verbs

An initial experiment was designed to determine whether a production network of the type shown in Figure 2 could learn some of the rules of Amharic verb formation. The training and test items for the network consisted of verb forms based on 30 triconsonantal roots. The forms varied by aspect (perfect, jussive, converb) and number-gender (third person singular masculine, third person singular feminine, third person plural). For the root *kft*, the nine possible forms are *kʌffʌtʌ* (3 s. m. perf.), *kʌffʌtʌčč* (3 s. f. perf.), *kʌffʌtu* (3 p. perf.), *kʌfto* (3 s. m. conv.), *kʌfta* (3 s. f. conv.), *kʌftʌw* (3 p. conv.), *yikfʌt* (3 s. m. juss.), *tikfʌt* (3 s. f. juss.), and *yikfʌtu* (3 p. juss.). For each root, six forms were selected randomly for the training set; the remaining three belonged to the test set.

A production network with 30 hidden units was used for this experiment. For each training sequence, the network was given a pattern representing the root, aspect, and number-gender as input and a sequence of segments as the output targets. Following training, errors were made on only 35 of the 664 output segments (5.3%; chance: 97.1%) in the test items. The network has clearly made the appropriate generalizations.

Tier-Like Representations in the Hidden Layer

The question now is, in learning the rules of Amharic verb formation, is the network extracting dimensions which correspond to the tiers of an autosegmental analysis? I will confine the investigation to a search for evidence of a root tier.

First a production network identical to the one used in the previous experiment was trained on all 270 sequences. Next the hidden layer patterns for all sequences (2010 patterns in all) were saved, and 10 principal components for these vectors were extracted. Finally the paths traced by hidden layer patterns for individual sequences along the components were examined.

On several of the components, component 1 in particular, there is a clear direction of movement from the first to the second root consonant. The distinction between root consonants 2 and 3 is not made so clearly, but there is a consistent pattern on component 9 between these two positions. Figure 3 shows values along component 1 plotted against those along component 9 for three forms based on the root *rgṭ*. The movement in this two-dimensional space is consistently rightward from the first to the second root consonant and downward from the second to the third root consonant regardless of where the consonants occur in the word. Note also that it is these directions which the different words share rather than the particular location of the points in the space. The characteristic movement is also independent of the particular consonants that make up the root. For example, Figure 4 shows the corresponding patterns for three words based on the root *ṭrg*, which consists of the same consonants as *rgṭ* but in different root positions.

Together principal components 1 and 9 constitute a space

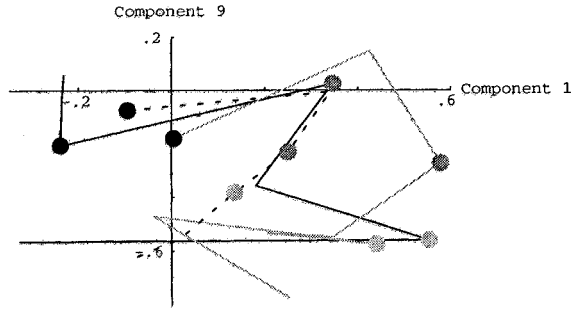

Figure 3: Paths along Components 1 and 9 for 3 Forms of *rgt*

Jussive (solid line), perfect (fuzzy line), converb (dashed line); Root consonants 1 (dark dots), 2 (medium dots), 3 (light dots)

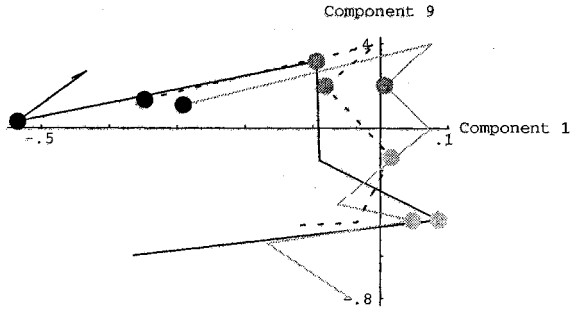

Figure 4: Paths along Components 1 and 9 for 3 Forms of *trg*

Jussive (solid line), perfect (fuzzy line), converb (dashed line); Root consonants 1 (dark dots), 2 (medium dots), 3 (light dots)

within which there is movement in a characteristic direction as the consonants of a root are produced in the output sequence. Because this movement depends neither on absolute position in the output sequence nor on the particular consonants making up the root, this two-dimensional space seems analogous to the root tier of an autosegmental analysis.

DISCUSSION

Where does phonology come from? One possibility is that the learner is pre-programmed somehow to look for regularity in the input, and this sort of learning probably does play a role. Another is that the learner acquires phonology as a side-effect of the process of lexical acquisition. Just as a linguist, in analyzing the phonological system of an unfamiliar language, looks for phonetic contrasts that distinguish morphemes, so the child might be expected to develop phonological abstractions on the basis of what makes a lexical/grammatical difference. The experiments described in this paper represent an initial attempt to demonstrate how this sort of phonological learning might take place in a system which is not designed specifically to learn phonology, but which learns it because it needs to in order to recognize or produce words.

This approach fits in with a growing body of work demonstrating that connectionist networks can make relatively sophisticated linguistic generalizations on the basis of large bodies of input (e.g., [1, 6, 7]). These approaches contrast with those which posit highly abstract built-in constraints on what can be learned (e.g., [2]). In connectionist models, what is innate is the architecture of the processing/learning system, and a fundamental question for linguistics and cognitive science concerns the extent to which architectural constraints compatible with connectionism are adequate for learning language. This paper lends some support to this possibility.

References

[1] D. P. Corina. *Towards an Understanding of the Syllable: Evidence from Linguistic, Psychological, and Connectionist Investigations of Syllable Structure.* PhD thesis, University of California, San Diego, 1991.

[2] B. E. Dresher and J. D. Kaye. A computational learning model for metrical phonology. *Cognition*, 34:137–195, 1990.

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

[4] L. Frazier. Structure in auditory word recognition. *Cognition*, 25:157–187, 1987.

[5] J. Goldsmith. *Autosegmental and Metrical Phonology.* Basil Blackwell, Cambridge, MA, 1990.

[6] Prahlad Gupta and David S. Touretzky. Connectionist networks and linguistic theory: Investigations of stress systems in language. Unpublished report, Carnegie-Mellon University, 1991.

[7] M. L. Hare. The role of similarity in Hungarian vowel harmony: a connectionist account. *Connection Science*, 2, 1990.

[8] M. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 531–546, Hillsdale, New Jersey, 1986. Lawrence Erlbaum Associates.

[9] W. J. M. Levelt. *Speaking: From Intention to Articulation.* MIT Press, Cambridge, MA, 1989.

[10] J. J. McCarthy. *Formal Problems in Semitic Phonology and Morphology.* Garland, New York, 1982.

[11] J. L. McClelland, D. E. Rumelhart, and G. E. Hinton. The appeal of Parallel Distributed Processing. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1*, pages 3–44. MIT Press, Cambridge, MA, 1986.

[12] R. Port. Representation and recognition of temporal patterns. *Connection Science*, 2:151–176, 1990.

[13] D. E. Rumelhart, G. E. Hinton, and R. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing, Volume 1*, pages 318–364. MIT Press, Cambridge, MA, 1986.

[14] T. van Gelder. Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, 14:355–384, 1990.