In B. MacWhinney (Ed.) (1999). The Emergence of Language. Mahwah, NJ: Lawrence Erlbaum

Chapter 16

The Emergence of Language From Serial Order and Procedural Memory

Prahlad Gupta
Gary S. Dell
Beckman Institute for Advanced Science & Technology
University of Illinois at Urbana–Champaign

In every spoken language, the words have two fundamental properties.[1] First, they are temporal sequences; in the articulatory domain, a word is a sequence of gestures, and in the auditory domain, a sequence of sounds. Second, the relation between this serially ordered sequence and the word's meaning is arbitrary. Words with similar meaning do not typically have similar forms.

It is a truism that the nature of language has shaped the way that it is learned and processed. It is also generally agreed that the reverse is true; learning and processing constraints have profoundly influenced language. In this chapter, we start with the fact that words are serially ordered temporal sequences arbitrarily related to meaning, and ask how these properties reflect and are reflected in the processing system. Specifically, in the first section, we suggest that some aspects of vocabulary structure can be linked to the fact that words are sequences. Along with this, we will consider the production of words and introduce a simple recurrent network model that relates some speech error phenomena to the sequential nature of words. This account of sequence and speech errors embodies a view in which many aspects of language learning can be understood as proceeding via the gradual, experience-driven adjustment of connection weights between levels of representation. In the second section, we examine parallels between this

[1]We are using the term *word* informally and intuitively. We are not referring to a specific constituent such as the phonological word.

general view of language learning and implicit or procedural memory. In particular, we suggest that the systematic aspects of language learning are continuous with the mechanisms of procedural memory and show that the simple recurrent network model that accounts for speech errors can also account for important phenomena in the domain of procedural memory. In the third section, we situate our speech error model in the cognitive system and situate our discussion of procedural memory more concretely in the context of language. We discuss how word forms can be acquired and associated with arbitrary meanings. We suggest that the seemingly unitary process of learning a new word can be fractionated into two components that rely on two quite different memory systems and learning mechanisms. The discussion centers on the relation between the sequential and arbitrary aspects of words and on the correspondence of these aspects of word learning with the procedural and declarative memory systems. Overall, we propose in this chapter that important aspects of language can be viewed as emerging from serial order and from procedural memory.

WORDS AS TEMPORAL SEQUENCES

Recently, *Scientific American* published a debate between. Stephen Hawking and Roger Penrose on the fate of the universe (Hawking & Penrose, 1996). The article started with an assumption that both debaters accepted: The "big bang," the rapid expansion at the beginning of the universe, and the "big crunch," when everything collapses together again at the end, will likely look very different. The big crunch will not be the big bang in reverse. However, if this is true, there is a puzzle. The laws of physics, as we know them, are symmetric in time. Consequently, there is no explanation for the assumed asymmetry on a cosmological scale. So, why will the end of the universe be different from the beginning? The article did not offer a definitive answer. Nonetheless, it was agreed that something must provide the universe with temporal asymmetry.

The debate in *Scientific American* leads into the first question that we ask: Why are the beginnings of words different from their ends? If we break words down into sequences of phonemes or features, we find several temporal asymmetries in the vocabulary. For example, languages have positional phonotactic constraints. Certain sounds (e.g., /ŋ/ or /ʒ/ in English) cannot begin words, whereas others (e.g., /h/ and many short vowels in English) cannot end them. Furthermore, the constraints on sound combinations differ at the beginning and ends of words (and syllables). The onset–rime organization of syllables is an example. Within a syllable there are many more constraints between the vowel and following (coda) consonants than between the vowel and preceding (onset) consonants. In

English CVC words, for example, the consonant /r/ can precede any vowel, but it cannot follow most vowels (Kessler & Treiman, 1997).

The differences between words' beginnings and ends require us to postulate some temporal asymmetry. Fortunately, in cognitive science, we are not burdened with a theory whose laws are indifferent to time's arrow. Instead, the production or recognition of a sequence is inherently asymmetric because of memory. We remember the past much better than we can predict the future. As we process a sequence we have a better representation of prior context, specifically of elements in the sequence that have already been processed, than we do of subsequent context. On the assumption that representations of context guide our processing, there is reason to expect differences between the beginnings and ends of words.

The sequential asymmetries in the lexicon are reflected in speech production errors. First, consider positional phonotactic constraints. Speech errors are exquisitely sensitive to these constraints. Slips in which a phoneme occurs in a phonotactically illegal position, such as *neck* spoken as *ngeck*, are rare in comparison to cases such as *meck* for *neck*, in which the wrong sound occupies a legal position. Stemberger (1983), for example, found that less than 1% of phonological errors were phonotactically deviant.

Another example of a speech error effect tied to sequence concerns the onset–rime asymmetry. The onset and rime constituents of syllables appear to act as units in errors. Consider the phrase *smell bad*, whose constituent structure is illustrated in Fig. 16.1. One is more likely to see the exchange of the rime constituents /ɛl/ and /æd/ (*smad bell*), or the onsets /sm/ and /b/ (*bell smad*), than exchanges of other contiguous groups of sounds such as /smɛ/ and /bæ/ (*bal smed*).

A final speech error phenomenon linked to sequence is the onset bias. Most phonological speech errors involve the movement or substitution of word-initial consonants. In English, for instance, around 80% of consonant exchanges involve word-initial sounds (Shattuck-Hufnagel, 1987). Although some languages (e.g., Spanish) do not exhibit a word-onset bias, their errors nonetheless have a strong syllable-onset bias (Berg, 1991)

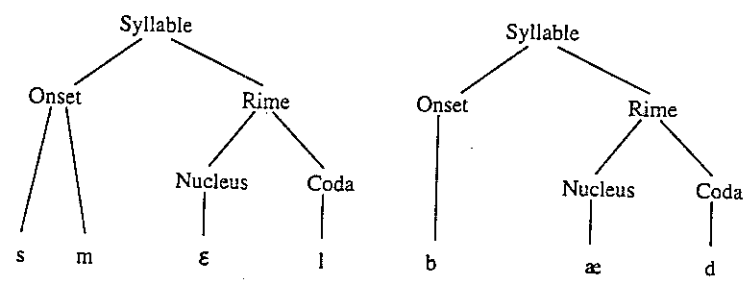

FIG. 16.1. Phonological constituent structure of the example phrase *smell bad*.

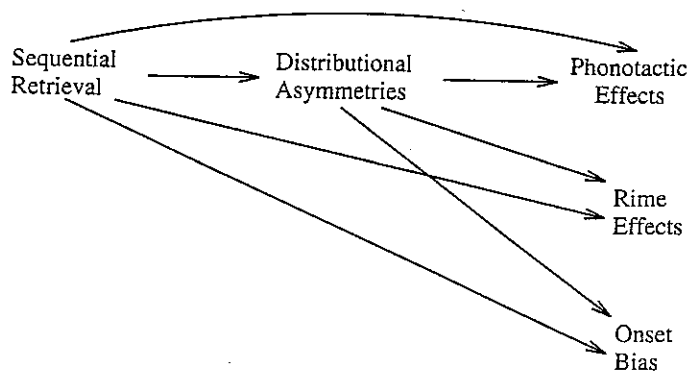

FIG. 16.2. Causal links hypothesized between the sequential nature of words, distributional asymmetries in the lexicon, and speech errors.

Figure 16.2 illustrates the causal links that we hypothesize between these speech error effects and the serially ordered nature of words. The causal structure has two parts. First, we hypothesize that the fact of temporal sequence leads to distributional asymmetries in the vocabulary. Specifically, we claim that the need to retrieve the sounds of words in sequence favors a vocabulary with asymmetries such as an onset–rime organization. Second, we argue that the speech error effects derive jointly from the asymmetries in the vocabulary and directly from the sequential nature of retrieval during production. In the next two subsections, we flesh out these claims, beginning with the link between serial order and the vocabulary.

Vocabulary Structure, Sequential Cueing, and Interference

The fact that words are temporal sequences has consequences for how they are perceived and produced. One consequence is the sequential cueing assumption: Sounds that have already been processed serve as cues that guide subsequent processing. So, after hearing or retrieving the first sounds of a word, the processing system is biased toward later sounds that are consistent with this prior context. Because it emphasizes the constraint imposed by prior context, the sequential cueing assumption is temporally asymmetric. Prior context is hypothesized to be a more powerful cue than subsequent context. We can most definitely see this asymmetry in word recognition, where decades of research have shown that the initial sounds of an acoustic string have a disproportionate impact on the candidate words that are actively considered (Marslen-Wilson & Zwitserlood, 1989).

Sequential cueing also may apply to sound retrieval during production. Sevald and Dell (1994) asked speakers to say word pairs as many times as they could during a 4-sec interval. The word pairs either shared two initial

sounds (CAT CAB), one initial sound (CAT CUB), one final sound (CAT BUT), or two final sounds (CAT BAT). The pairs that were similar at the beginning were much harder to recite than those that were similar at the end. The effects were stronger when two sounds were shared rather than one. Figure 16.3 shows the difference in the production time per syllable for the pairs sharing two initial sounds and the pairs sharing two final

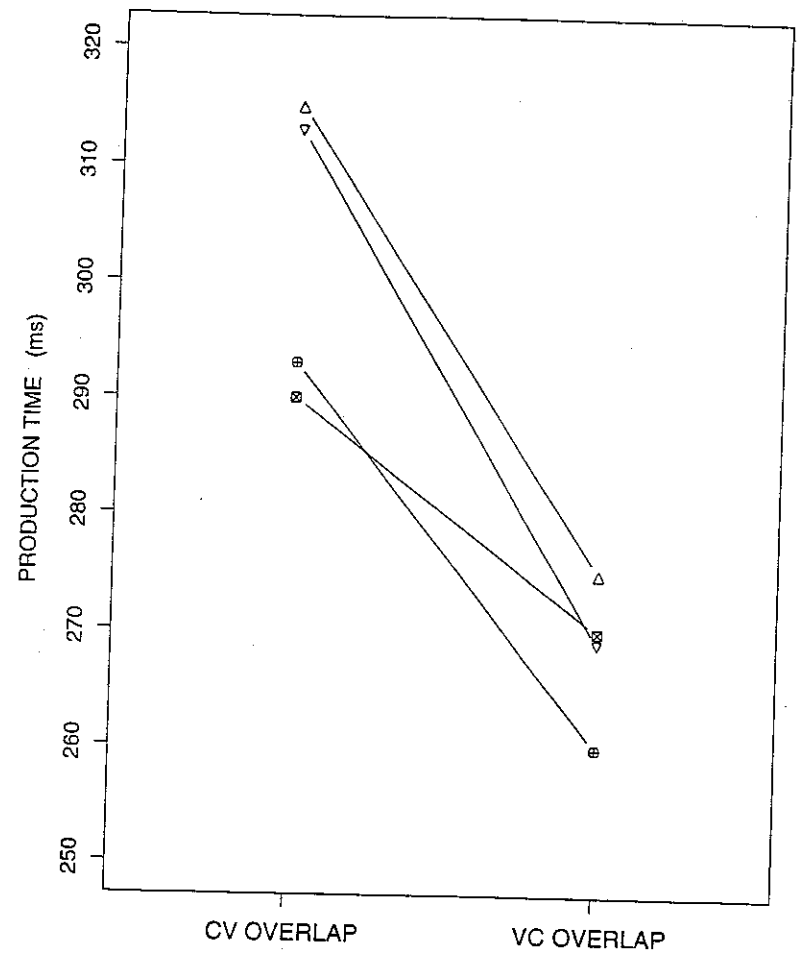

FIG. 16.3. Production time per syllable in repetition of pairs of CVC words that are the same either at the beginning (CV overlap) or at the end (VC overlap). The four lines plot the difference for four different groups of word pairs. Within each group, the word pairs were constructed from a particular set of consonants and vowels. Redrawn from Sevald and Dell (1994).

sounds. The 40-msec difference is quite large considering that the average production time for a syllable is less than 300 msec.

The difficulty that speakers have in reciting word sequences with common initial sounds can be called the *sequential interference effect.* It reflects, according to Sevald and Dell (1994), the greater cueing power of prior over subsequent context. When a speaker is seeking the final sound of the word CAT, for example, the already retrieved /kæ/ acts as a retrieval cue. However, when saying CAT CAB CAT CAB etc., this cue is equally associated with /t/ and /b/ and hence is a source of interference, rather than being a benefit. Interference is much less in evidence, however, when the word pair is similar at the end, as in CAT BAT. Here, the distinguishing sounds, /k/ and /b/, are followed rather than preceded by identical contexts. There is less interference because subsequent within-word context is not represented as strongly as prior context.

The repeated recitation of a word pair can be conceptualized as rapid retrieval from a minilexicon consisting of two words. The results of Sevald and Dell (1994) then suggest that it is easier to produce words from a minilexicon of CAT and BAT than one with CAT and CAB. A related point can be made about word recognition. A CAT-BAT lexicon would be associated with faster recognition than a CAT-CAB one. Generally speaking, on the assumption that words do share sounds, it is best that the common sounds be at the end. Applying these observations to language-sized lexicons suggests that some vocabularies are better than others. A good vocabulary would contain word sets such as CAT, BAT, SAT, RAT, MAT, and so on, whereas a bad one would have sets like CAT, CAN, CAB, CAD, and CAP. To put it another way, a good lexicon's words are not randomly distributed in phonological space. Instead, words should have neighbors sharing sounds at the end rather than neighbors sharing sounds at the beginning.

Real lexicons appear to be organized in exactly this way. Kessler and Treiman (1997) analyzed the sound distribution in the English CVC lexicon, focusing on differences between the ends of words—the VCs—and the beginnings—the CVs. The set of VCs is quite restricted. Several possible VCs, such as /Un/ or /iŋ/, do not occur at all, whereas others such as /æt/ or /ab/ are much more likely than would be expected from the frequency of their constituent sounds. The CVs, in contrast, are largely unrestricted; just about any initial consonant can be followed by any vowel. Moreover, the frequency of a particular CV is predictable from the frequency of the consonant and vowel. A consequence of these facts is that, on average, CVC words have more VC neighbors than CV neighbors.

We suggest that the structure of the vocabulary has been shaped by the sequential nature of words, in particular by the interference that results when competing words share initial sounds. In traditional phonological

theory, the distributional asymmetries noted by Kessler and Treiman have been explained by the assumption that syllables have an onset–rime hierarchical organization. Because the VC of a CVC word is a single constituent, there are more restrictions between the vowel and the coda than between the vowel and the onset. These restrictions are expressed by cooccurrence rules that allow or prohibit particular sound combinations in the rime. Our view is different. The onset–rime organization in the vocabulary emerges from sequential factors. So, we would explain the distributional facts by appealing to the processing consequences of sequentiality, rather than by assuming an onset–rime structure. Specifically, we suggest that forces on vocabulary evolution such as the need for rapid retrieval favor redundancy toward the ends of words, leading to neighborhood structures in which a selected set of endings (rimes) recur.[2]

An important aspect of Kessler and Treiman's (1997) analysis favors an account of distributional asymmetries in terms of sequential factors instead of an assumption that syllables possess an onset–rime organization. Kessler and Treiman found that the restrictions on VCs were graded rather than absolute. In addition to nonoccurring VCs such as /Un/, there are several VCs whose frequency in the vocabulary is much less than would be expected. For instance, /ʌl/ occurs in significantly fewer words than would be expected. (See also Frisch, 1996, for other evidence for graded phonotactic constraints and for the influence of sequential factors on these constraints). One might explain the graded nature of the constraints by proposing that the cooccurrence rules only indirectly reflect a universal a priori onset–rime organization. However, it is more parsimonious, in our view, to dispense with the a priori and treat onset–rime structure as emergent from sequential processing. The mechanism is simply that more difficult words, for example, those whose neighbors have common beginnings as opposed to common endings, are less likely to be used and, hence, will tend to drop out across generations. This would presumably be a gradual probabilistic process and so one would expect to see the graded cooccurrence restrictions that Kessler and Treiman found.

Our linkage of distributional asymmetries in the lexicon to factors such as sequential interference is speculative. We have no direct evidence for the hypothesized evolutionary mechanism. However, what is not speculative is the fact that the statistical structure of the word-form lexicon influences perception and production. Sound combinations that are common in the

[2]To make this account really work, we would need a mechanism for the sound distributions in single-syllable words to transfer to the syllables of multisyllabic words, because onset–rime organization is about syllables rather than words. PDP models of the production and recognition of phonological sequences (e.g., Plaut & Kello, chap. 14, this volume; Dell et al., 1993) may have such a mechanism based on shared weights, but this remains to be seen because the models have not yet been applied to multisyllabic words.

lexicon are easy to perceive and produce (Levelt & Wheeldon, 1994). In the following section, we turn to the second half of the causal chain presented in Fig. 16.2 and show how the distributional asymmetries can explain the speech error effects.

The Speech-Error Model

The model that we review here (Dell, Juliano, & Govindjee, 1993) is a simple recurrent network model of word-form retrieval. For our purposes, it has three key properties: it retrieves words as sequences of phonological features; it uses prior context to guide the retrieval of the sequence (and hence embodies the sequential cueing assumption); and it is sensitive to the statistical structure of the word-form lexicon through its learning mechanism.

Figure 16.4 shows the architecture. There are four layers of units: input, hidden, output, and context. The input layer represents the word to be spoken as a distributed pattern of activation (1s and 0s). The output consists of a bank of units, one for each of 18 phonological features. Sequences of features are produced by means of recurrent one-to-one connections from the output and hidden layers to the context layer. There were two sets of context units, internal context units, whose activations were copies of the hidden layer's activation on the previous pass through the network, and external context units, whose activations derive analogously from the output units. The context units thus serve as a dynamic memory for prior

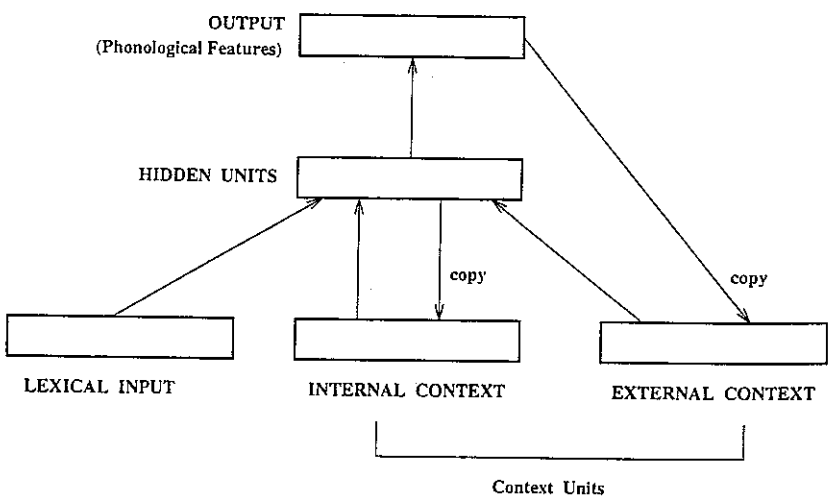

FIG. 16.4. Simple recurrent network model of word-form retrieval and speech errors (redrawn from Dell et al., 1993).

context and act as input along with the static input layer to cue for the appropriate features throughout the production of the word.

Specifically, the production of a word involves the following steps, illustrated here using CAT. The input pattern for CAT is entered on the input units, the internal context units are initialized to zero, and the external context units are set to .5, a pattern that symbolizes a word boundary. Activation then spreads to the output features via the hidden units. The target output is the feature set for the first phoneme, here /k/. To the extent that the output deviates from the target, weights are adjusted by backpropagation (Rumelhart, Hinton, & Williams, 1986). Then the output and hidden layer activations are copied to the external and internal context units, respectively. Hence, after the output of /k/, the context units represent the state of /k/ just having been retrieved, instead of their original state that marked the beginning of the word. This change in the context then allows for the production of /æ/ in the next forward pass through the network. The process continues until the end of the word, the final target being the word boundary pattern. The model was trained by repeatedly presenting short English words (1–3 phonemes), and adjusting the weights until performance was mostly correct.

The modeling goals were to explain certain speech error facts. These included the phonotactic regularity of errors, the tendency for multiphoneme errors to coincide with syllabic constituents such as the rime, and the tendency for errors to occur at the beginnings of words. Earlier, we suggested that these three error effects are related to the sequential nature of words and distributional asymmetries in the lexicon resulting from sequence.

If noise is introduced to the connection weights, the model produces realistic phonological errors, and these errors exhibit the three effects in question. First, the erroneous sequences obey the phonotactic constraints of English. Versions of the model that were trained on representative samples of English words made errors that were phonotactically legal between 90% and 100% of the time. For example, the model might produce *gat* or *cot* instead of *cat* but rarely *ngat* or *ctt*. Second, the model exhibits sensitivity to the rime constituent. Its multiple-phoneme substitution errors are about two or three times more likely to involve rime (VC) constituents (e.g. *bed* slips to *big*) than CV combinations (e.g., bed slips to *mad*). Finally, the model's errors showed the onset effect. Errors involving single consonants involved word onsets more often than other consonants (between 59% and 78% were onsets) even though there were slightly more nononset than onset consonants in the training vocabularies.

Why does the model produce these speech error effects? In general, the model's errors are sensitive to the structure of English words because it is trained on English words. The superimposed weight changes associated with

the training set create sequential schemata, pathways in activation space that reflect common sequences of features. When the model makes an error it tends to stick close to those pathways, or move to other well-worn paths in the vicinity. That is why the errors obey phonotactic constraints. The reason that the model's errors sometimes involve rime units is also due to the structure of the training set. When that set mirrors English in having a restricted set of rimes each of which is part of several words, the occurring rimes then become part of the well-worn paths in the model. When output jumps from its target path to another, that path is likely to be one that is frequently taken, that is, one corresponding to a common rime.

The tendency for errors to obey phonotactic constraints and to coincide with rimes arises in the model simply because the model responds directly to the statistics of its training set. Any distributional asymmetries in the training set are reflected in the model's behavior. Hence, these two error phenomena relate to the sequential nature of words only indirectly, through properties of the vocabulary. The third error effect, the onset effect, comes directly from the model's sequential properties. Because the model uses prior context as input, the first sound is more difficult to retrieve than the others. Consequently, there are more errors involving onset consonants. At the beginning of a word, the context units' activation is completely uninformative about the features to be retrieved. The lexical input must be relied on entirely to signal the features of the first phoneme. However, as more and more of the word is retrieved, the context units become more helpful. There is simply greater constraint later on in the word than at the beginning.

In summary, the speech-error model allows us to see how error effects that are associated with the sequential structure of words arise. We claim that any model in this domain must, first, be sensitive to the statistics of the sound patterns in the vocabulary, including patterns that are temporally asymmetric, and second, the model must itself produce output in sequence with an emphasis on the cues provided by prior context. Simple recurrent networks, in general, have both of these properties.

To conclude our discussion of speech errors and sequence, we turn to a new experimental finding that highlights people's abilities to learn sound distributions and how those distributions affect errors.

Learning Sequences of Sounds

As noted previously, the reason why speech errors adhere to phonotactic constraints is, in our view, that these errors reflect experience with the statistical patterns in the vocabulary. However, patterns occur at many levels of generality. For example, one can speak of patterns that are true for the entire language. The facts that /h/ is an onset or /ŋ/ is a coda

are true for all of English (and many other languages, for that matter.) There are also patterns that apply only to a subset of words, such as the location of stressed syllables in English verbs. Ultimately, one can speak of "patterns" that are so specific that they apply to single words. For example, CAT is a CVC word with two voiceless stop consonants: /k/ as onset, /t/ as coda, and /æ/ as the vowel.

One can view some speech error effects in terms of very local patterns. The most common phonological speech errors involve the movement of sounds from one syllable or word to another, such as *leading list* for *reading list* or *black bloxes* for *black boxes* (Fromkin, 1971). Typically, these movement errors exhibit what is called the syllable-position effect: When a sound moves from one place to another, it retains its position in the syllable.[3] So, in *leading list* the erroneous /l/ in *leading* is an onset just like its source in *list*.

We suggest that the syllable position effect and the fact that errors obey positional phonotactic constraints are related phenomena. They just refer to constraints at different levels of generality. Consider the fact that /h/ is always a syllable onset when one is speaking English. This is a language-wide constraint, and errors adhere to it, as seen in the absence of any errors with /h/ as a coda. Next, consider the fact that /k/ is always an onset when one is saying the word CAT. This is a local constraint, one that is true only for CAT. An error in the vicinity of CAT, involving the production of /k/, will nonetheless preserve the onset status of /k/. For example, *the cat chases* might be spoken as *the cat cases*. The migrating /k/ preserves the position it is constrained to occupy in its originating word. In this way, the syllable position effect can be thought of as adherence to a local constraint in the vicinity of that word. Adherence to language-wide positional phonotactic constraints is simply a more general application of this principle: Substituting phonemes obey the constraints on positions in which they can occur in the language.

The syllable position and phonotactic effects can be thought of as two ends of a continuum of breadth of constraint. The syllable position effect is at the narrow end and adherence to phonotactics is at the wide end. The often discussed distinction between rules and exceptions is analogous. A rule applies widely, and exceptions locally. So, the past tense for English is *ed*, but the past tense of the word *come* is *came*. Moreover, the regularities in English past tense morphology fall on a continuum between rules and exceptions and include a variety of subregularities such as the *ring–rang, sing–sang, drink–drank* cluster.

If breadth of constraint is truly a continuum, we ought to be able to find a speech error effect that is in the "middle" between the syllable-

[3]Sometimes the constraint is formulated in terms of word position; see Meyer (1992).

position effect and error adherence to language-wide phonotactics. An experiment by Dell, Adams, Reed, and Meyer (1998) provided exactly this. They showed that speech errors in an experimental paradigm exhibit the syllable-position effect, language-wide phonotactic effects, and a new effect, sensitivity to learned experiment-wide constraints on the position of sounds within the syllable.

The experimental participants were given a simple task: They had to recite four-syllable lists three times in a row in time with a metronome. They did this for four 1-hour sessions, each of which was associated with the recitation of 96 different lists. Each syllable in the lists was a CVC and included the vowel /ɛ/. The eight consonants in each list always included exactly one of each of the following: /h/, /ŋ/, /f/, /s/, /n/, /m/, /k/, and /g/. For example, one string was *hes feg meng ken.*

There were three sets of constraints on the location of the eight consonants in each list, language-wide, experiment-wide, and list-wide constraints. The language-wide constraints concerned /h/ and /ŋ/. Because /h/ is an onset and /ŋ/ is a coda in English, these constraints were present in the experimental lists as well. The experiment-wide constraints concerned /f/ and /s/. For each experimental participant, the syllable position of these two sounds was held constant across the experiment. Half of the participants only experienced /f/ as an onset and /s/ as a coda, and half only experienced the reverse. These experiment-wide constraints thus represented a level of constraint that is less general than the language-wide constraints. The relevant context is "syllables in this experiment" as opposed to "syllables in English." The list-wide constraints applied to the other four sounds, /n/, /m/, /k/, and /g/. Because each sound occurs exactly once in a list, each of these four sounds is experienced as either an onset or a coda for this list. However, the next list may change the sound's syllable position. The list-wide constraints are therefore the least general, the relevant context being "syllables in this list."

The effect of the three constraints was assessed simply by determining whether the speakers' errors during the recitation of the lists obeyed or violated the constraint. As expected, the language-wide constraints were upheld without exception. There were 498 cases in which the erroneously spoken sound was /h/ or /ŋ/ (e.g., *meng ken* → *meng keng*) and in every case, the sound appeared in a phonotactically legal position. Thus, this provides an experimental demonstration of the phonotactic legality of errors.

Next, consider the list-wide constraints. Here, the errors involved the production of /n/, /m/, /k/, or /g/ instead of a target sound. These errors could obey or violate list-wide constraints. For example, *meng ken* spoken as *meng men* upholds the constraint that /m/ is an onset for this list. If these syllables are spoken instead as *meng kem*, the error violates the constraint that /m/ is an onset. The results showed an imperfect adherence

to the constraint: 873 illegal to 323 illegal errors, 73% adherence. Thus, list-wide constraints are not upheld to the extent that language-wide constraints are, which is what one would expect from a system that responds directly to the strength of constraint. Notice that this result constitutes an experimental demonstration of the syllable-position effect: A moving sound tends to retain its syllable position. However, we are conceptualizing this effect in terms of a constraint of narrow breadth.

The key result of the experiment concerns experiment-wide constraints. That is, do the errors in which /f/ and /s/ replace target sounds obey the experiment-wide constraints on the position of these sounds? If there is zero sensitivity to the experiment-wide constraint, then we would expect the errors involving /f/ and /s/ to be legal at about the same rate as those testing for the sequence-wide constraint (73%). This is because the experiment-wide constraint implies sequence-wide constraint. On the other hand, if language-wide constraints reflect nothing more than recent experience with the distribution of the sounds /h/ and /ŋ/ in the experiment, then one would expect the experiment-wide results to be as exceptionless as the language-wide ones.

It turned out that the language-wide constraints were very strongly upheld, but were not exceptionless. There were 276 relevant legal errors (e.g., *meng* spoken as *feng* for participants for whom /f/ was an onset) and only 3 illegal ones (e.g., *kem* spoken as *kef*). This level of adherence—nearly 99%—suggests a highly adaptive production system, one that is constantly updating its knowledge of the position of sounds within syllables. Moreover, the fact that the experiment-wide comparison fell between the language-wide and sequence-wide ones tells us that breadth of constraint is, in fact, graded. The constraints with wider applicability are more likely to be upheld in errors.

Another aspect of the results with the experiment-wide constraint is that they illustrate implicit learning. When one erroneously produces an /f/, one is clearly not deliberately accessing the knowledge about whether /f/ is an onset or a coda. There is no time for such deliberation and, moreover, what was produced was an error rather than the intended sound. Further evidence for the implicit nature of the learning was provided by two other findings. First, the experimental participants reported no awareness of the distribution of the key sounds /f/ and /s/ at the end of the experiment. Second, half of the participants were told the rule about the distribution of /f/ and /s/ at the beginning of the experiment; their error data did not differ from the participants who were not informed.

In summary, the experiment tells us that the production system responds to experience. Error adherence to sound patterns, and perhaps other manifestations of knowledge about sound patterns, are products of implicit learning from experience with the sequential patterns in the vocabulary

and may, in fact, be largely due to recent experience. It is worth noting that these conclusions are quite in line with the simple recurrent network model that we have reviewed here. If learning is always assumed to be "on," and there is a component of the input that reflects the situation (e.g., being in a particular experiment), the model's weights will come to be sensitive to the sound patterns in that situation.

So far in the chapter, we have discussed how onset–rime distributional asymmetries in the lexicon may arise from the needs of sequential processing. We also showed that some speech-error effects derive jointly from the asymmetries and other distributional statistics in the vocabulary and directly from the sequential nature of retrieval during production. We described experimental results showing that many phonological constraints in language are the result of learning contingencies at multiple levels of temporal structure. In the rest of this chapter, we argue that these kinds of phenomena are really manifestations of domain-general procedural learning mechanisms, deployed within the specific domain of language. We show that similar effects arise in nonlinguistic sequential domains and can be accounted for by essentially the same incremental learning mechanisms as described earlier. We focus particularly on how sensitivity to multiple levels of structure directly reflects experience with the distributional statistics of the vocabulary in nonlinguistic domains, just as in linguistic domains.

LANGUAGE LEARNING AND PROCEDURAL MEMORY

The foregoing discussion of speech errors and sound-sequence learning incorporates a view of language learning that has been increasingly adopted in recent years. According to this view, many aspects of language learning can be understood as proceeding via the gradual, experience-driven adjustment of connection weights between levels of distributed representation. This view has been applied to aspects of language as varied as sentence comprehension (McClelland, St. John, & Taraban, 1989), inflectional morphology (Gupta & MacWhinney, 1992; Hoeffner, 1992; MacWhinney & Leinbach, 1991; Rumelhart & McClelland, 1987), phonology (Dell et al., 1993; Gupta & Touretzky, 1994), and reading (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989), and it is also incorporated in many of the models described in this volume. This is, of course, the view of learning and cognitive processing embodied in the Parallel Distributed Processing (PDP) framework (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986).

These same kinds of learning mechanisms have also, however, been proposed for another domain of cognition, namely, implicit or procedural memory. One of the most significant developments in the study of human

memory over the last 2 decades has been the discovery of a dissociation between two different kinds of memory systems, referred to as explicit versus implicit memory (Schacter, 1987) or declarative versus procedural memory (Cohen, 1984; Cohen & Squire, 1980). Dissociations between procedural and declarative memory were originally discovered in amnesics (Cohen & Squire, 1980) and have since also been documented in normal populations (e.g., Graf & Schacter, 1985). Declarative memory refers to memory for arbitrarily related collections of representations, such as episodes, or factual knowledge. Implicit or procedural memory includes the acquisition of skills that are acquired gradually over several sessions of practice and facilitation or priming in processing of a stimulus following prior exposure to that stimulus (e.g., Squire, 1992). Much of the interest in procedural or implicit memory arises from the fact that a great deal of everyday human learning appears to have just this character, occurring gradually as a result of practice over many exposures. This is, of course, precisely the view of language learning described previously. In fact, it has been hypothesized (e.g., Cohen, 1984; Cohen & Eichenbaum, 1993) that procedural memory is based on the continual, experience-driven tuning of processing elements, similar to the experience-driven adjustment of connection weights in PDP networks. Thus, within the domain of memory research, one major human memory system, procedural memory, has been characterized as operating via the same principles that are incorporated in many recent models of language. A similar view of procedural memory is inherent in the work of some of the architects of the PDP framework (e.g., McClelland, McNaughton, & O'Reilly, 1995; McClelland & Rumelhart, 1985).

Bringing these ideas together, we hypothesize that those aspects of language learning that can be characterized as gradual, experience-driven tuning are in fact forms of procedural memory, a suggestion that has also been made by a number of other researchers (e.g., Cleeremans, 1993; Reber, 1993). This suggestion extends the domain of procedural memory to encompass many of the systematic aspects of language learning and implies that these aspects of language learning can be viewed as *emerging* from the mechanisms of procedural memory.

If this view is correct, it should be possible to use the same kinds of computational models that have been proposed for language, to account for phenomena in the domain of procedural memory. In what follows, we discuss how a model very similar to the model of speech errors described earlier (Dell et al., 1993) can yield insight into aspects of procedural memory. In our earlier discussion of this model, the focus was on the temporal aspects of processing, manifested in such phenomena as speech errors. We now focus instead on how such a system might develop. Let us suppose that the model is exposed to a number of word forms in each period of time. In each period of time (such as a day), some of these word forms

are ones that have previously been encountered, whereas some of the word forms are novel, as depicted in Fig. 16.5a. Let us further suppose that the model's task is simply to spell out each input word form representation, as a sequence of output phonemes or phonological features (see Fig. 16.4). How might the system's basic ability to perform this task develop? Figure 16.5b shows the pattern of results that we might expect. The ability to spell out familiar forms would improve greatly, as a result of multiple exposures to these forms; this is depicted by the lower curve in Fig. 16.5b. Additionally, we would expect cumulative practice to result in generalized performance improvement, as a result of continual adjustment of connection weights in the system. That is, we would expect the model to become more effective at the process of spelling out word forms in general, irrespective of whether they were familiar or not. This is shown in the upper curve in Fig. 16.5b.

This examination of the effects of repetition on novel and familiar stimuli is precisely what is studied under the general heading of "skill learning" and "repetition priming" in the implicit memory literature. Figure 16.5c schematizes an experimental paradigm that is commonly used in studies of repetition priming (e.g., Kirsner & Speelman, 1996; Logan, 1990; Schwartz & Hashtroudi, 1991). In this paradigm, participants are required to respond in some way to each stimulus with which they are presented. Each block of stimuli is comprised of a number of "unique" stimuli that appear exactly once throughout the experiment and a number of "repeating" stimuli that appear in every block. Figure 16.5d illustrates the phenomena of interest in such studies: improvement in performance on unique stimuli (termed *skill learning*) and the greater facilitation of performance on repeating stimuli over performance on unique stimuli (this difference being *repetition priming*).

There appears to be a clear analogy between the developmental effects we would expect in our PDP model of speech errors and the study of skill learning and repetition priming as aspects of procedural memory. In the remainder of this section we examine this analogy further, by discussing skill learning and repetition priming in a specific procedural memory task, with a view to making two points. First, we show that learning occurs at multiple levels of structure in this procedural memory task, just as in the sound-sequence learning experiment of Dell et al. (1998) described earlier. In the sound-sequence task, we observed sensitivity to distributional statistics at the level of experiment-wide constraints, list-wide constraints, and language-wide constraints. We describe a procedural memory task in which, similarly, sensitivity to sequential constraints arises at multiple levels. Second, we show that a simple recurrent network essentially identical to the Dell et al. (1993) model can account for these and other results from the procedural memory task. These two demonstrations emphasize the conti-

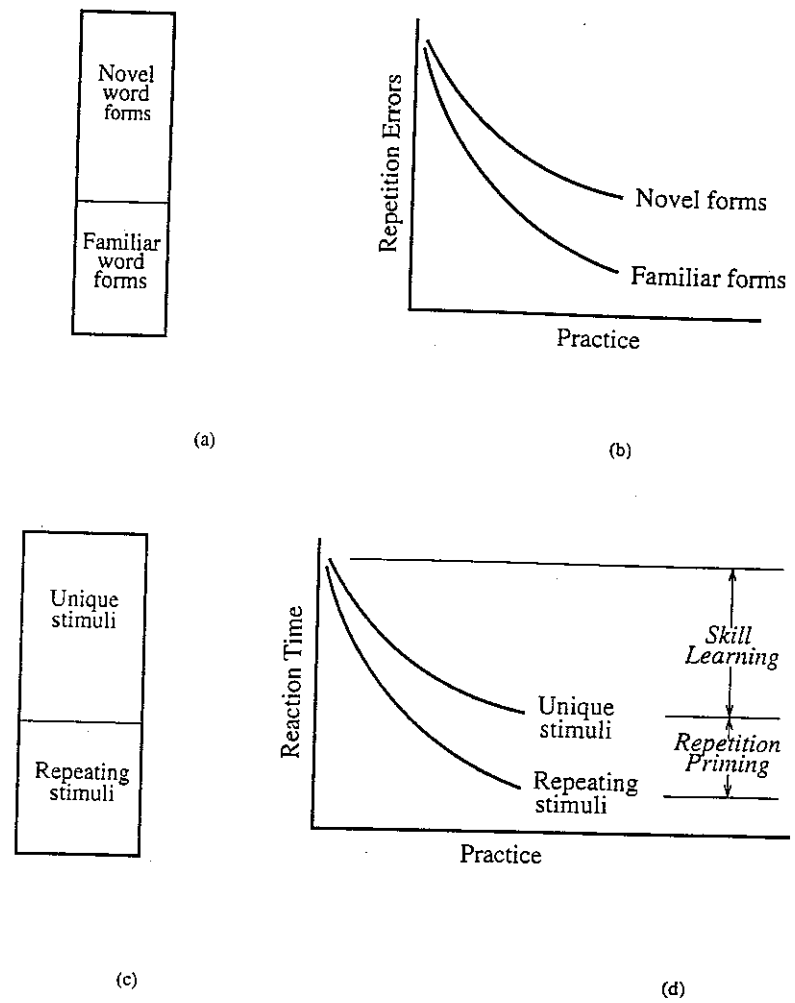

FIG. 16.5. Comparison of expected development of the ability to repeat word forms, in a sequential PDP model, with issues in the study of two procedural memory phenomena, skill learning and repetition priming: (a) hypothetical regimen of environmental exposure for the PDP system; (b) expected schedule of development; (c) regimen of experimental practice in a typical procedural memory task; (d) measures of interest in such procedural memory paradigms.

nuity of the phenomena and mechanisms of procedural memory with those of language learning.

Skill Learning and Repetition Priming

We focus on the digit entering task (Fendrich, Healy, & Bourne, 1991), in which both skill learning and repetition priming have been studied. We describe a version in which 5-digit numbers (e.g., *49385*) were presented individually to participants, who entered these numbers on a computer numeric keypad (Poldrack, 1995, Experiment 1a). No feedback was provided for the response. After entering the number, the participant pressed the enter key, which triggered the next trial after a 1-sec interval. Participants were instructed not to correct errors. Some numbers appeared multiple times during the experiment (repeating items) whereas other items appeared only once (unique items).

Training stimuli were chosen from the set of all possible 5-digit numbers by placing three constraints on them. First, only the digits 1 through 9, were allowed (i.e., the digit 0 was not used). Second, digits could not repeat immediately (e.g., *44* never appeared as part of a number). Third, 5-digit numbers were constrained to obey certain first-order transitions between digits, as follows. For each of the digits 1 through 9, there are 8 possible (different) digits that can follow it. For each digit, 4 of these 8 possible transitions were chosen at random. The set of 36 chosen transitions (4 transitions for each of the 9 digits) comprised one full transition rule set. There was also a complementary transition rule set comprised of the other 36 transitions (4 complementary transitions for each of the 9 digits). Figure 16.6 illustrates such transition rule sets for the digit 1. Five-digit numbers were constrained to obey one or other of these transition rule sets. Half of the participants in each experiment were trained on 5-digit numbers that followed one transition rule set, and the other half were trained on items following the complementary rule set.

The experiment (Poldrack, 1995, Experiment 1a) consisted of three sessions, each consisting of 12 blocks. The structure of the experiment is summarized in Table 16.1. Forty-eight 5-digit strings were presented in each block. However, the composition of the 48 stimuli in a block varied across sessions. In Session 1, each block consisted of 12 repeating stimuli, that is, stimuli that appeared once in every block throughout Session 1 and also throughout the experiment ("S1 Repeats"). The remaining 36 stimuli in each block of Session 1 were unique stimuli, that is, stimuli that appeared once, in only that one block and that did not appear in any other block in the experiment ("S1 Uniques"). In Session 2, each block consisted of the 12 repeating stimuli from Session 1; 12 new repeating stimuli that appeared in each block in Session 2 and through the rest of

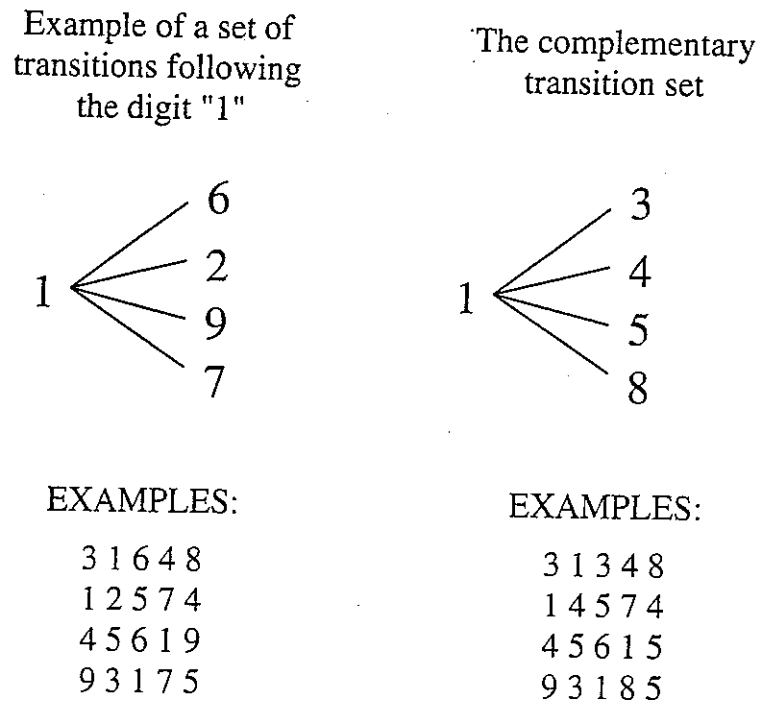

FIG. 16.6. Example of transition rules used in the digit entering task of Poldrack (1995, Experiment 1a).

TABLE 16.1
Structure of the Digit Entering Task of Poldrack (1995, Experiment 1a)

	Trial Type	*Number of Stimuli per Block*
Session 1	S1 Repeats	12
	S1 Uniques	36
Session 2	S1 Repeats	12
	S2 Repeats	12
	S2 Uniques	24
Session 3	S1 Repeats	12
	S2 Repeats	12
	S3 Uniques	12
	New Rule Uniques	12

the experiment ("S2 Repeats"), and 24 unique stimuli ("S2 Uniques"). In Session 3, each block consisted of the 12 Session 1 repeating stimuli, the 12 Session 2 repeating stimuli, and 12 unique stimuli ("S3 Uniques"). In addition, each block in Session 3 contained a further 12 unique stimuli that did not conform to the transition rule structure of all the other stimuli ("New Rule Uniques"). That is, all the stimuli in all three sessions of the experiment conformed to a particular set of transition rules, except for the Session 3 New Rule Uniques, which conformed to the complementary rule set.

Response times were separated into two components, latency for the first keystroke, and interkeystroke interval. Our concern is with interkeystroke interval (IKI), the average interval between each keystroke subsequent to the first. Figure 16.7 illustrates pertinent results from Poldrack (1995, Experiment 1a). The horizontal axis shows blocks of practice whereas the vertical axis shows participants' performance as measured by mean IKI per block. The four curves in the upper part of the figure show (from the top downward) performance on S3 New Rule Unique stimuli, performance on Old Rule Unique stimuli, performance on S2 Repeat stimuli, and performance on S1 Repeat stimuli. Skill learning was measured as the improvement in performance on Old Rule Uniques over the course of the experiment. That is, skill learning at block *n* is the difference between performance on Old Rule Uniques at Block 1 and performance on Old Rule Uniques at Block *n*. This measure is plotted as the upper of the two curves in the lower part of the figure. Repetition priming at Block *n* was measured as the difference between performance on Old Rule Unique stimuli and performance on S1 Repeat stimuli, both measured at Block *n*. This measure is plotted as the lower of the two curves in the lower part of the figure.

The following effects are noteworthy. First, there is marked improvement in performance on both Repeating stimuli and Old Rule Unique stimuli across blocks in Fig. 16.7. Second, there is a clear increase in skill learning across blocks, seen in the decrease in IKI for Old Rule Unique items across blocks and also seen in the plot of skill learning in the lower part of the figure. There clearly is also repetition priming, as shown by the difference between performance on Old Rule Unique stimuli and performance on S1 Repeat stimuli at each block and as also shown by the plot of repetition priming in the lower part of the figure. Third, repetition priming can be observed at different levels of skill learning: In addition to repetition priming for the S1 Repeats introduced at the beginning of the experiment, there is also repetition priming for S2 Repeats, which were introduced only in Session 2. The latter effect can be seen in the divergence in performance between S2 Repeat stimuli and Old Rule Unique stimuli, following introduction of the S2 Repeat stimuli in Block 13. Fourth, there is negative transfer to the New Rule Unique stimuli introduced in Session

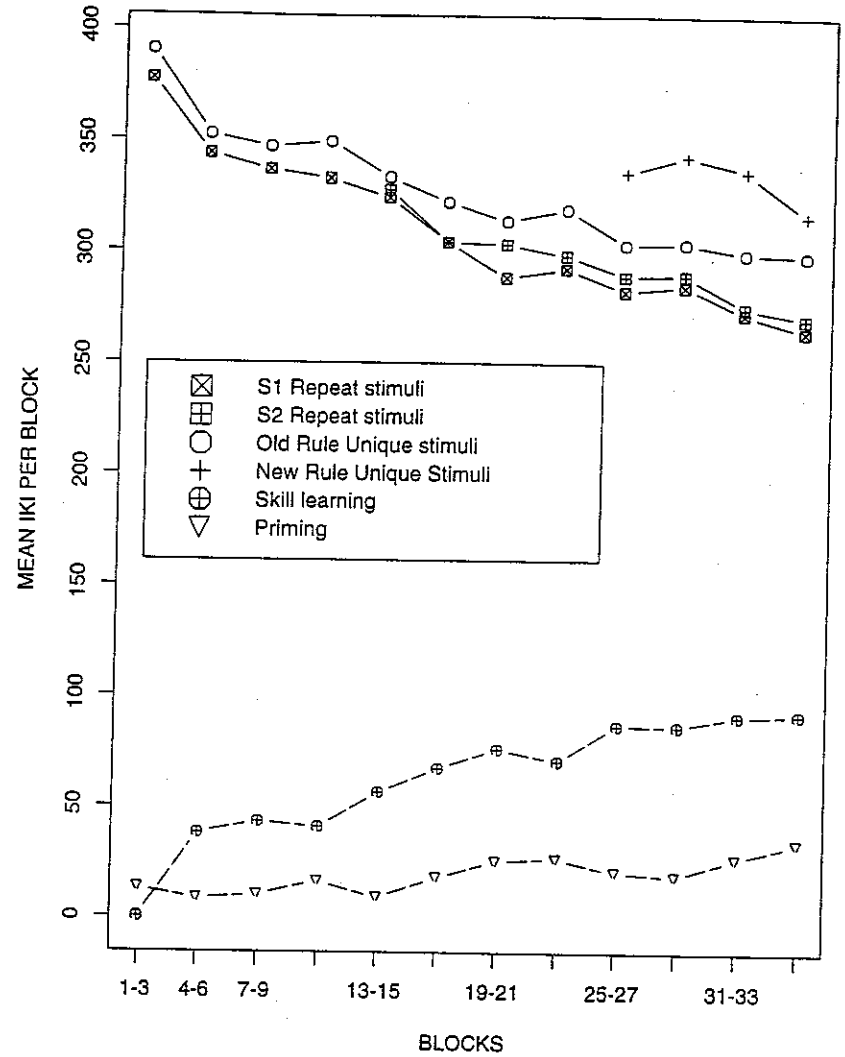

FIG. 16.7. Key results redrawn from the digit entering task of Poldrack (1995, Experiment 1a).

3. The only difference between the Old Rule Unique stimuli and New Rule Unique stimuli is in the familiarity of the specific 2-digit transitions they incorporate. The negative transfer effect therefore indicates that participants learned something about the 2-digit transition structure of the Old Rule Unique stimuli. That is, part of what they learned were subitem regularities within the stimuli.

For present purposes, one of the most significant aspects of these results is that participants appear to learn structure at multiple levels, just as in the sound-sequence learning experiment of Dell et al. (1998) described earlier in this chapter. In that experiment, the pattern of speech errors revealed participants' sensitivity to within-experiment regularities as well as to within-list regularities. These multiple levels of sensitivity necessarily must have developed during the course of the experiment. In Poldrack's (1995) results, participants learned regularities at both the 5-digit level and the 2-digit level during the course of the experiment. Sensitivity to the former is shown by superior performance on Repeating over Unique stimuli. Sensitivity to the latter is shown by superior performance on Old Rule Unique stimuli over New Rule Unique stimuli, in Session 3.

Figure 16.8 illustrates the architecture of the model that Gupta and Cohen (1997) used to simulate the digit entering task. The model is a simple recurrent network (Cleeremans, Servan-Schreiber, & McClelland, 1989; Elman, 1990; Jordan, 1986) and is essentially the same as the model of Dell et al. (1993). The task for the present model was formulated to have the same structure as the digit entering task. In each trial in the digit entering task, participants were presented with a 5-digit number, which they were required to convert into a sequence of keyboard strokes representing the sequence of digits comprising the 5-digit number. Similarly, the input to the model was a representation of a 5-digit number, and the model's task was to produce a sequence of outputs representing the sequence of digits in the input.

The input was represented as a pattern of activation over five banks of units that constituted the input layer, as shown in Fig. 16.8. Each of these

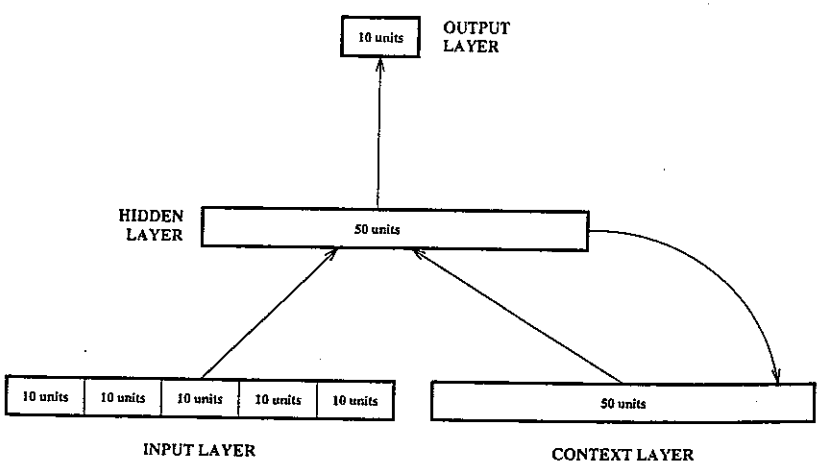

FIG. 16.8. Architecture of the model of the digit entering task (Gupta & Cohen, 1997).

banks comprised 10 units representing the digits 0 through 9, with each unit representing a specific digit. Thus, the 5-digit string *49683* was represented as a pattern of activation in which the 5th unit in the first bank of units was active (representing the occurrence of *4* as the 1st digit in the string); the 10th unit in the second bank of units was active, representing the occurrence of *9* as the 2nd digit in the string; the 7th unit in the third bank was active, representing the occurrence of *6* in third position; the 9th unit in the fourth bank was active, representing the occurrence of *9* as the 4th digit; and the 4th unit in the fifth bank was active, representing the occurrence of *3* as the 5th digit. All other units in the 50-unit input layer were inactive. The output layer consisted of a single bank of 10 units, representing the digits 0 through 9, with each output unit representing a specific digit. When presented with an input pattern of the kind described previously, the network's task was to spell out the input as a sequence of activations at the output layer. For example, for the input *49683*, the network's task was to activate, in sequence, the output layer units representing the digits *4*, *9*, *6*, *8*, and *3*.

At the first step in processing this input, activations from the input layer and from the context layer were propagated forward to the hidden layer, and activations from the hidden layer in turn were propagated forward to the output layer. This resulted in some pattern of activation at the output layer. An error signal was generated, representing the distance of this actual activation pattern from the target pattern, in which the output unit representing *4* should be maximally activated and all other output units inactive. Back propagation of error (Rumelhart et al., 1986) was then used to adjust the strength of connection weights throughout the system.

At the next step in processing the input, the activation of hidden layer units was copied to the context layer (Elman, 1990), and the network was now expected to produce as output the 2nd digit in the input string. Note that the input pattern representing the digit string *49683* was still present. As on the first step of processing, the actual output produced by the network was compared with the target output for the 2nd digit, and connection weight strengths adjusted so as to reduce the magnitude of error. Processing continued in this way for three further time steps, at which the network's task was to output the 3rd, 4th, and 5th digits, respectively, of the input. At the end of the fifth time step, the context layer was cleared, and the next 5-digit input string was presented for the network to spell out.

The model was used to simulate results from the digit entering task of Poldrack (1995, Experiment 1a) summarized earlier. Sixteen participants took part in that experiment. One complete simulated "replication," accordingly, comprised 16 runs of the model. Each run of the model consisted of three parts. First, weights in the system were initialized to random values, corresponding to variation between participants. Second, the model

was trained on 1,000 5-digit strings, which were randomly chosen from the set of all possible 5-digit strings but excluded strings that would appear in the actual experimental simulations. This training consisted of four *epochs*, i.e., four cycles of presentation of the 1,000-stimulus set. The model's task at each 5-digit stimulus presentation was to spell out the sequence of digits comprising the string. This second step of pretraining was intended to provide the model with a preexperimental level of skill approximately equivalent to that of participants entering the experiment. At the end of preexperimental training, the model was correctly able to spell out 85.3% of novel 5-digit inputs, which corresponded quite well with the 86.98% accuracy of human participants on unique stimuli in the first block of the digit entering task.

The third step consisted of the actual experimental phase of the simulation run. The structure of this third phase precisely mirrored the structure of the digit entering task (Poldrack, 1995, Experiment 1a), which was summarized in Table 16.1. Thus, in the first epoch of a simulation, the model was presented with 48 stimuli (5-digit strings) that obeyed a particular transition rule structure. Twelve of these stimuli were to repeat throughout the simulation (S1 Repeats), whereas the other 36 stimuli would not be presented again during the simulation (S1 Uniques). As explained previously, the model's task at each 5-digit stimulus presentation was to spell out the sequence of digits comprising the string, just as in the experiment with human participants. Thus, Epoch 1 in the simulation corresponded to Block 1 in the Experiment. In the second epoch of a simulation, the 12 S1 Repeat stimuli were presented again, and 36 novel stimuli were also presented. Thus, Epoch 2 corresponded to Block 2. In analogous fashion, each epoch in the simulation mirrored the corresponding block of the experiment. It may be worth noting that the model's task in response to presentation of a 5-digit string was the same during Step 2 (pretraining) and Step 3 (experimental simulation). The only difference was that the presentation of stimuli during Step 3 had a carefully controlled structure, identical to that in the digit entering task (Poldrack, 1995, Experiment 1a).

In the digit entering task, the specific stimuli in each category (S1 Repeats, S1 Uniques, etc.) were carefully counterbalanced across the 16 participants (Poldrack, 1995, Experiment 1a). In simulations, the model was run once with each of these actual sets of stimuli. Thus, one simulation run consisted of: (a) random weight initialization, (b) pretraining for 4 epochs, and (c) simulation of the 36 Blocks of Experiment 1a, using one of the 16 stimulus sets actually presented to participants. A set of 16 such runs constituted one simulated replication of Experiment 1a. Twenty such replications were run and the results averaged. Simulation results were measured by mean squared error per digit, which was the mean squared difference between the target output pattern and the actual output pattern

for every digit. For example, performance on unique stimuli in Epoch 1 of the simulations was measured by the squared error per digit, averaged over all digits presented in Epoch 1, averaged over all 16 simulated participants, averaged over all 20 replications. This measure was found to be linearly related to the IKI measure used in the behavioral experiments.

Figure 16.9a redisplays the results from Poldrack (1995, Experiment 1a). Figure 16.9b shows the corresponding simulation results (Gupta & Cohen, 1997). As can be seen, the simulations provide a good fit to the data and exhibit all of the phenomena characteristic of the empirical results. First, there is improvement in performance on both unique and repeating stimuli across blocks. Second, we see clear indication of both skill learning and repetition priming. Third, the effects of repetition can be observed even after significant skill learning has occurred, as shown by performance on the new repeating items (S2 Repeats) introduced in Session 2. Fourth, there is negative transfer to stimuli that follow the complementary transition rule structure, in Session 3. These results confirm that the kinds of incremental learning mechanisms widely employed as accounts of language learning can also account for procedural memory phenomena.

LEXICAL LEARNING AND PROCEDURAL AND DECLARATIVE MEMORY

How can we concretize the notion that aspects of language learning are forms of procedural memory? In this section, we present a simplified view of the lexical system in language. We use this framework to situate our model of speech errors within a broader context and to situate our discussion of procedural memory more firmly within a linguistic setting.

Our view of the lexical system is depicted in Fig. 16.10. According to this view, when the sequence of speech segments ("input phonemes") constituting a word is presented as input to the lexical system, this sequence of inputs leads to activation of an internal representation of the word form. We assume that this internal representation is phonologically structured and that it is distributed over many processing units. The mapping from sequences of input sound segments to internal representations is the "input phonology" mapping. The internal word form representation, once activated, can lead to the playing out of a sequence of "output phonemes," corresponding to the process of saying (producing) the word. The mapping from internal word form representations to sequences of output phonemes is the "output phonology" mapping. In addition to participating in the input and output phonology mappings, internal word form representations also have associative links to and from the level of semantic representations. Thus, when a word form representation is activated, it evokes activation of any semantic representation that may be associated with the word form.

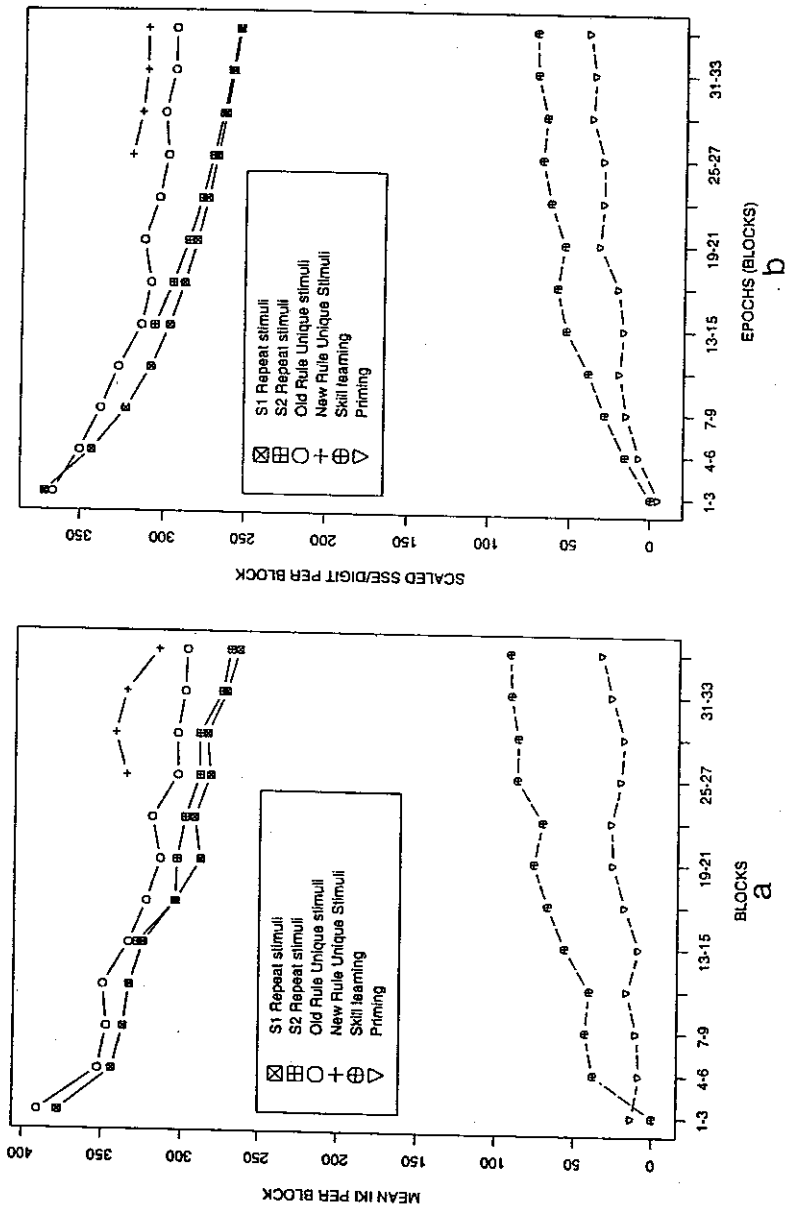

FIG. 16.9. Comparison of experimental and simulation results: (a) results from the digit entering task (Poldrack, 1995, Experiment 1a); (b) simulation results (Gupta & Cohen, 1997).

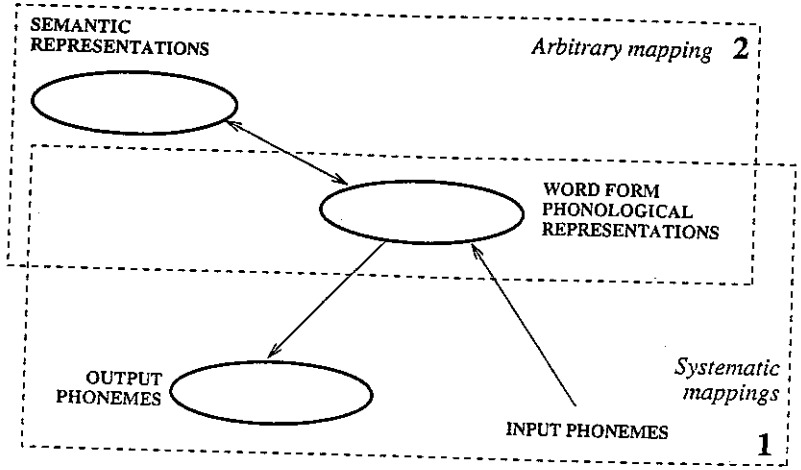

FIG. 16.10. One view of the lexical system in language and the nature of the mappings involved.

Conversely, when a semantic representation is activated, it leads to activation of any associated word form level representation.

The output phonology mapping is implemented in the cognitive system by a transducer that learns to approximate the mapping through experience-driven, incremental tuning. The kind of transducer we envisage is the system of weighted connections that mediates between two levels of representation in a PDP network. The incremental tuning is the kind of adjustment of connection weights that occurs in PDP networks. Similarly, the input phonology mapping and the two mappings between internal word form representations and semantics are also implemented by transducers that are incrementally tuned by experience.

The system we outlined is comprised of systematic as well as arbitrary mappings, as shown in Fig. 16.10. The input and output phonology mappings are both systematic, in that similar sequences of input phonemes map onto similar internal word form representations, and similar word form representations map onto similar sequences of output phonemes. However, as we noted at the beginning of this chapter, it is a fundamental property of all spoken languages that the mapping between word forms and meanings is arbitrary, in that similar word form representations are not guaranteed to map onto similar meanings, and vice versa. For example, the similar-sounding word forms *carlin*, *car*, and *marlin* do not map onto similar meanings. Conversely, the similar semantics of the objects "cup" and "glass" do not map onto similar-sounding word forms. Of course, in most languages, there are partial regularities in the mapping. For example, inflectional morphology leads to certain aspects of meaning (e.g., plural number) being expressed

consistently in a particular fashion (by adding *-s*, for example, in English). Nevertheless, the mapping is for the most part arbitrary.

How does the simple recurrent network model of speech errors described earlier fit in with the view of the lexical system we have just presented? Figure 16.11 illustrates the correspondence. In Fig. 16.11, a simplified version of the Dell et al. (1993) model is shown inverted, with its input layer at the top and its output layer at the bottom. Its input layer corresponds to the word form level of representation in the present lexical system whereas its output layer corresponds to the present output phoneme layer. The model additionally has a recurrent hidden layer not depicted in Fig. 16.10. Comparison of Fig. 16.11 with Fig. 16.10 should make it clear that the Dell et al. (1993) model is in effect an implementation of the present output phonology mapping. It instantiates one of the systematic mappings in the lexical system and is precisely the kind of transducer we envisaged.

The import of our observations about systematic versus arbitrary mappings becomes clearer when we consider the fact that learning a new word can in general be a fast process. Research on "fast mapping" indicates that children as young as 4 years of age can learn the association between a novel word form and its referent, given only one or two exposures to the pairing (Carey, 1978; Dollaghan, 1985), and certainly, adults have the ability to learn new words with very little repetition. Learning a new word requires establishment of a new entry in each of the transducers in the lexical system. Therefore, if the word is to be learned within one or two exposures, then this requires that new entries be established in each of the transducers with only one or two exposures.

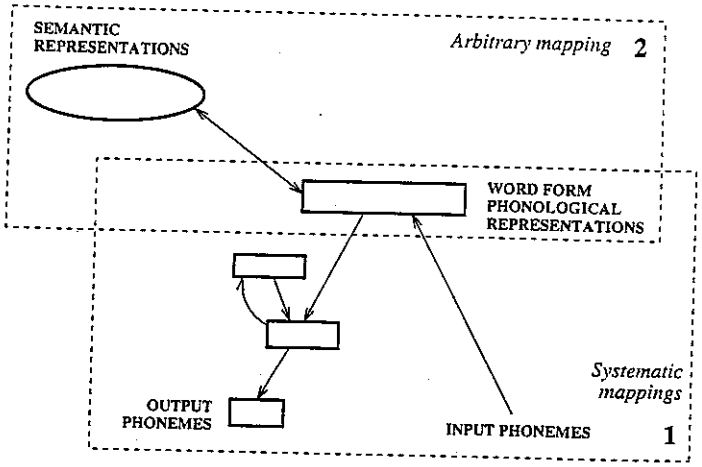

FIG. 16.11. Situating the Dell et al. (1993) model of speech errors within the lexical system.

Let us first consider the transducer that implements the output phonology mapping. Given that word form level representations are distributed, and given that the output phonology mapping is systematic, it follows that this transducer can generalize. It automatically has the ability to transduce a novel word form representation into the correct sequence of output phonemes. Thus, in the output phonology mapping, fast learning comes for free, as a process of similarity-based generalization. The same is true of the input phonology mapping. To see this, consider the human ability to repeat nonwords. Nonword repetition requires mapping from the novel input phoneme sequence to a novel word form level representation via the input phonology transducer, and then spelling out the novel word form representation as a novel sequence of output phonemes, via the output phonology transducer. The fact that humans can easily repeat nonwords shows that there must be generalization in both the input and output phonology mappings. This is not true, however, of the mapping from word forms to meaning, or from meaning to word forms, because these are arbitrary mappings. Similarity-based generalizations cannot determine the correct word form level representation for a novel semantic representation, or vice versa (although it can support productive inflectional processes, as previously noted). Furthermore, the connection weight changes needed to establish a new entry in these two transducers can be large. Gradual, incremental tuning processes will not suffice to make such large weight changes within one or two exposures. However, if we try to make the large necessary weight changes nongradually, previous entries in the transducers will be overwritten, given that the representations are distributed. Some other learning mechanism is needed for the mappings between word forms and semantics, to reconcile the conflicting requirements of fast learning and distributed representations.

To reiterate, the incremental, experience-driven tuning of connection weights provides a sufficient basis for fast learning in the systematic input and output phonology mappings. However, some additional mechanism is needed for fast learning in the arbitrary mappings between word forms and semantics. This distinction is reminiscent of the previously mentioned dissociation between explicit versus implicit or declarative versus procedural memory. Recall that declarative memory refers to memory for arbitrarily related collections of representations. For example, the set of representations comprising the memory of a past event or "episode" are forms of declarative memory (e.g., Squire, 1992), as is the factual knowledge that the capital of Madagascar is Antananarivo. Procedural memory includes the acquisition of skills that are acquired gradually over several sessions of practice, and facilitation or priming in processing of a stimulus, following prior exposure to that stimulus (e.g., Squire, 1992). There is considerable agreement about the neural basis of the dissociation between declarative

and procedural memory. The declarative memory system is believed to be subserved by the hippocampus and related medial temporal lobe structures. These structures provide for the initial encoding of memories involving arbitrary conjunctions and also for their eventual consolidation and storage in neocortex (e.g., Cohen & Squire, 1980; Mishkin, Malamut, & Bachevalier, 1984; Squire, Knowlton, & Musen, 1993). The procedural memory system, which provides for the learning and processing of motor, perceptual, and cognitive skills, is believed to be subserved by learning that occurs in neocortex and in the basal ganglia (e.g., Cohen & Squire, 1980; Mishkin et al., 1984; Squire et al., 1993).

It has further been suggested that there is a functional reason for the existence of two memory systems (McClelland et al., 1995). According to this view, the neocortex processes distributed representations. Such representations have the desirable property of providing a basis for similarity-based generalization. However, in a computational system employing distributed representations, learning of arbitrary associations must be slow and interleaved, if the system is not to suffer from "catastrophic interference," the overwriting of previously existing knowledge (McClelland et al., 1995; McCloskey & Cohen, 1989). Episodes and new facts (i.e., the kinds of arbitrary associations that constitute declarative memory) therefore cannot be encoded swiftly via slow cortical learning. It has been proposed that the role of the hippocampus is to convert distributed representations into localist nonoverlapping ones and to establish fast mappings between such converted representations. That is, the hippocampus performs fast learning, based on orthogonalized representations, and thus provides a basis for the swift encoding of arbitrary associations of the kind that comprise episodic and factual information (McClelland et al., 1995). Neocortex and hippocampus thus perform complementary learning functions and provide the basis for procedural and declarative memory, respectively.

It thus appears that the two different kinds of mechanisms required in lexical learning map quite well onto the functional roles assigned to procedural and declarative memory. This suggests that the input and output phonology mappings in our lexical system may be subserved by procedural learning, but that the mappings between word forms and meaning must invoke the mechanisms of declarative memory. To elaborate, the locus of the various phonological and semantic representations can be viewed as being cortical tissue (Blumstein, 1995; Gupta & MacWhinney, 1997; Kandel & Schwartz, 1985). Incremental tuning of the weighted connections between these cortical representations occurs automatically, following exposure to each conjunction of representations. This tuning occurs in all the transducers shown in Fig. 16.11—the input phonology transducer, the output phonology transducer, and the transducers between word forms and semantics. Although such tuning provides a sufficient basis for generalization and fast learning in the input and output phonology transducers, it cannot support fast learning in the transducers between word forms and semantics. It needs to be augmented by hippocampally mediated learning processes that are not represented in Fig. 16.11 but that are added in Fig. 16.12. The cortical representations are relayed to the hippocampus where they are bound together rapidly in the form of orthogonalized representations. In the long term, however, the binding of semantic and word form representations to each other must occur in the connections between the cortical representations, and this process of consolidation proceeds slowly because cortical learning is slow (Cohen & Eichenbaum, 1993; McClelland et al., 1995). Once established cortically, a particular binding can be thought of as having become proceduralized.

Two predictions follow from the view we outlined. First, under conditions of impairment in declarative memory, it should be difficult to learn new word meanings quickly. Second, even if declarative memory is impaired, it should nevertheless be possible for procedural learning to occur in the input and output phonology transductions; we might expect such tuning to be manifested in the form of repetition priming. These are in fact precisely the kinds of impairments observed in hippocampal amnesics. Such patients are virtually unable to learn new word meanings (e.g., Gabrieli, Cohen, & Corkin, 1988; Grossman, 1987), which is an indication of their impairment in declarative memory. However, these same patients exhibit intact repetition priming for both known and novel words (e.g., Haist, Musen, & Squire, 1991), which is an indication of their relatively spared procedural memory. These results support the hypothesis that phonological learning proceeds via cortical procedural learning mecha-

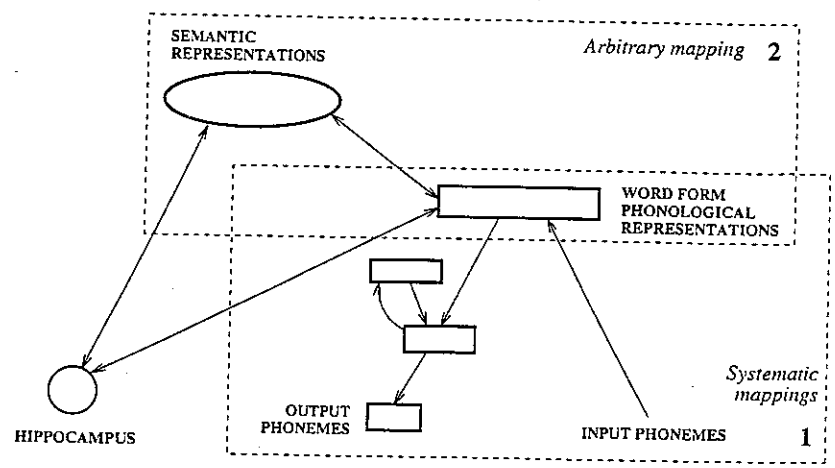

FIG. 16.12. Hypothesized role of the hippocampus in lexical learning.

nisms, whereas learning the mapping between word forms and meaning relies additionally on hippocampal declarative memory.

CONCLUSIONS

We began this chapter with the observation that in all spoken languages words are temporal sequences arbitrarily related to meaning, and we asked how these properties reflect and are reflected in the processing system. We suggested that the processing mechanisms necessary to deal with the serially ordered nature of words play a role in the emergence of distributional asymmetries in the lexicon. Specifically, we claimed that the need to retrieve the sounds of words in sequence favors a vocabulary with asymmetries such as an onset-rime organization. We then argued that some speech error effects derive jointly from the asymmetries and other distributional statistics in the vocabulary and directly from the sequential nature of retrieval during production. We reviewed a simple recurrent network model that implements such an account of speech error phenomena. We also presented new experimental results supporting the view that many phonological constraints in language are the result of learning contingencies at multiple levels of temporal structure.

We then suggested that those aspects of language learning that can be characterized in terms of the slow, incremental tuning of connection weights may in fact be forms of procedural memory. We focused particularly on how sensitivity to sequential constraints can arise at multiple levels in procedural memory tasks, similar to the sensitivity to multiple levels of constraint observed in our sound-sequence learning task. We described the application of a simple recurrent network (similar to the speech-error model) to data from a procedural memory paradigm, and showed that the model provides an excellent account of the empirical results. The similarity in effects and the applicability of the kinds of incremental learning mechanisms proposed for language processing to the domain of procedural memory emphasizes the continuity of the phenomena and mechanisms of procedural memory with those of language learning.

Finally, we situated our speech error model in the lexical system and provided a specific example of the role of procedural memory in language learning. We argued that the existence of different kinds of mappings in the lexical system leads to a pressure for two different mechanisms to support fast learning. We suggested that these mechanisms can be identified with those of procedural and declarative memory and showed that the available neuropsychological data are consistent with this hypothesis. The seemingly unitary process of learning a new word thus appears to draw on two rather different systems of learning and memory.

Overall, our discussion in this chapter suggests that important aspects of language can be viewed as emerging from serial order and from procedural memory, and that lexical learning draws on the mechanisms of procedural as well as declarative memory.

ACKNOWLEDGMENTS

This work was supported by a Beckman Institute Fellowship to the first author, and by NSF Grant SBR 93-19368 and NIH Grants DC-00191 and HD-21011. We are grateful for the advice of David Adams, Kathryn Bock, Barbara Church, Neal Cohen, Victor Ferreira, Zenzi Griffin, Brian MacWhinney, Antje Meyer, Kevin Miller, Gregory Murphy, and Kristopher Reed. Correspondence and requests for reprints may be addressed to either author, at the Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801.

REFERENCES

Berg, T. (1991). Phonological processing in a syllable-timed language with pre-final stress: Evidence from Spanish speech error data. *Language and Cognitive Processes, 6,* 265–301.

Blumstein, S. E. (1995). The neurobiology of the sound structure of language. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 915–929). Cambridge, MA: MIT Press.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing.* Cambridge, MA: MIT Press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. (1989). Finite state automata and simple recurrent networks. *Neural Computation, 1,* 372–381.

Cohen, N. J. (1984). Preserved learning capacity in amnesia: Evidence for multiple memory systems. In L. R. Squire & N. Butters (Eds.), *Neuropsychology of memory* (pp. 83–103). New York: Guilford.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science, 210,* 207–209.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17,* 149–195.

Dell, G. S., Reed, K. D., Adams, D. A., & Meyer, A. S. (1998). *Speech errors, phonotactic constraints, and recent experience.* Unpublished manuscript.

Dollaghan, C. (1985). Child meets word: "Fast mapping" in preschool children. *Journal of Speech and Hearing Research, 28,* 449–454.

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

Fendrich, D. W., Healy, A. F, & Bourne, L. E. (1991). Long-term repetition effects for motoric and perceptual procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 137–151.

Frisch, S. (1996). *Similarity and frequency in phonology.* Doctoral dissertation, Northwestern University, Linguistics Department, Evanston, IL.

Fromkin, V. (1971). The nonanomalous nature of anomalous utterances. *Language, 47,* 27–52.

Gabrieli, J. D. E., Cohen, N. J., & Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain, 7,* 157–177.

Graf, P., & Schacter, D. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 501–518.

Grossman, M. (1987). Lexical acquisition in alcoholic Korsakoff psychosis. *Cortex, 23,* 631–644.

Gupta, P., & Cohen, N. J. (1997). *Skill learning, repetition priming, and procedural memory: Theoretical and computational analysis.* Manuscript submitted for publication.

Gupta, P., & MacWhinney, B. (1992). Integrating category acquisition with inflectional marking: A model of the German nominal system. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 253–258). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59,* 267–333.

Gupta, P., & Touretzky, D. S. (1994). Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science, 18,* 1–50.

Haist, F., Musen, G., & Squire, L. R. (1991). Intact priming of words and nonwords in amnesia. *Psychobiology, 19,* 275–285.

Hawking, S. W., & Penrose, R. (1996). The nature of space and time. *Scientific American, 275*(1), 60–65.

Hoeffner, J. (1992). Are rules a thing of the past? Learning the past tense in an attractor network. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 861–866). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach* (Report 8604). La Jolla: Institute for Cognitive Science, University of California, San Diego.

Kandel, E. R., & Schwartz, J. H. (Eds.). (1985). *Principles of neural science* (2nd ed.). New York: Elsevier.

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language, 37,* 295–311.

Kirsner, K., & Speelman, C. (1996). Skill acquisition and repetition priming: One principle, many processes? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 563–575.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50,* 239–269.

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology, 22,* 1–35.

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

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 576–585.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing, Vol. 2: Psychological and biological models.* Cambridge, MA: MIT Press.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4,* 287–335.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–165). New York: Academic Press.

Meyer, A. S. (1992). Investigation of phonological encoding through speech-error analyses: Achievements, limitations, and alternatives. *Cognition, 42,* 181–211.

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: Two neural systems. In G. Lynch, J. McGaugh, & N. Weinberger (Eds.), *Neurobiology, of learning and memory* (pp. 65–77). New York: Guilford.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Poldrack, R. A. (1995). *The relationship between skill learning and repetition priming.* Unpublished doctoral dissertation, Department of Psychology, University of Illinois, Urbana-Champaign.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious.* New York: Oxford University Press.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing, Vol. 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing, Vol. 1: Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processes? In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 195–248). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schacter, D. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518.

Schwartz, B. L., & Hashtroudi, S. (1991). Priming is independent of skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 1177–1187.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Sevald, C. A., & Dell, G. S. (1994). The sequential cueing effect in speech production. *Cognition, 53,* 91–127.

Shattuck-Hufnagel, S. (1987). The role of word-onset consonants in speech production planning: New evidence from speech errors. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (pp. 17–51). Hillsdale, NJ: Lawrence Erlbaum Associates.

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting memory and learning. *Journal of Cognitive Neuroscience, 4,* 232–243.

Squire, L. R., Knowlton, B., & Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology, 44,* 453–495.

Stemberger, J. P. (1983). *Speech errors and theoretical phonology: A review.* Bloomington: Indiana University Linguistics Club.