7

Neural Network Models of Speech Production

Matthew Goldrick

The ability to translate thoughts into articulatory gestures – resulting in sounds that others can perceive – is a critical part of human behavior. Concepts from neural or connectionist networks have played an important role in the development of theories in this domain. Following Smolensky (1999), this chapter focuses on three core principles of connectionist theories of speech production:

- *Mental representations are distributed, graded patterns of activation.* Mental representations are realized as gradient numerical patterns of activation over simple processing units.
- *Cognitive processing is the spread of activation.* Cognitive processes are realized by transformations of activity patterns by numerical connections.
- *Cognitive processing reflects the statistical structure of the environment.* The structure of cognitive processes reflects the ongoing modification of the spread of activation based on the statistics of the environment.

Following a brief introduction to connectionist networks, the impact of each of these principles on psycholinguistic theories of speech production is reviewed. The next section then examines how learning-based modification of spreading activation may (or may not) lead to novel theories of cognitive processes.

At the outset, it should be noted that connectionist networks are best thought of not as networks of realistic models of neurons but rather as neurally *inspired* networks. Although the principles of connectionist computation are broadly consistent with what is known regarding neuronal computational principles, connectionist processing mechanisms represent considerable abstractions from actual

The Handbook of the Neuropsychology of Language, First Edition. Edited by Miriam Faust.

neurobiological mechanisms. Connectionist researchers see this as a virtue; such networks provide an appropriate theoretical vocabulary for bridging cognitive- and neural-level explanations (Smolensky, 2006). This chapter aims to illustrate how the integration of cognitive concepts with neuronal processing principles has served to enhance theory development in the domain of speech production.

A Connectionist Primer

A connectionist network is a computational device composed of simple processing units linked by numerical connections or weights. The processing units are associated with a numerical quantity referred to as activation. Processing occurs via the propagation of activation along the numerical connections. This allows activation from one unit to "flow" to others, allowing one unit to influence the activation of others.

In many domains, cognitive processes can be conceptualized as input–output mappings – a relation between two sets of mental representations. For example, Palmer and Kimchi (1986, p. 40) define the components of cognitive psychological theories as "informational events" consisting of "the *input information* (what it starts with), the *operation* performed on the input (what gets done to the input) and the *output information* (what it ends up with) [emphasis original]." In a connectionist network realizing or modeling a cognitive process, mental representations are instantiated via patterns of activation over sets of units. The input to the cognitive process is realized by imposing an activation pattern over one set of units (the input units). Activation then spreads via connections to other units (instantiating the operation performed on the input). This results in a pattern of activation over a set of output units; the mental representation corresponding to this pattern of activity is the output.

To provide a concrete illustration of this framework, consider the simple network illustrated in Figure 7.1. This implements a cognitive process that maps conceptual representations to lexical representations. These representations are realized via patterns of activation over two sets of processing units. This example follows many

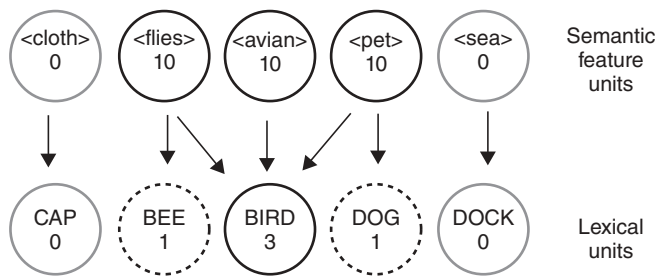

Figure 7.1 Illustration of the spread of distributed patterns of activation. All connections have strength 0.1. Activations are shown below unit labels.

connectionist theories of speech production (e.g., Dell, 1986) by making use of localist representations. In strictly local representations, each distinct mental representation is realized by a single processing unit (e.g., each word has an independent unit such as <BIRD> [angle brackets (< , >) denote the content of the mental representation instantiated by the unit]). In feature-based or semi-local representations a set of processing units realizes each distinct mental representation (e.g., each concept is encoded by a set of semantic features such as <avian>). In contrast, in a distributed representational framework there is not a strict relationship between elements of mental representations and processing units; individual processing units participate in the realization of many distinct mental representations (see Vousden, Brown, & Harley, 2000, for an example within phonological processing in speech production).

In this example, the concept "a flying avian pet" is provided as input to this cognitive process by activating the semantic feature units corresponding to this concept. Here they are assigned the arbitrary activation level of 10 (other features remain inactive). Activation then spreads along the connections between semantic features and lexical items. The amount of activation flowing to each lexical item is simply the sum over all incoming connections of the connections' weight times the activation of the sending unit. For example, the amount of activation flowing to <BEE> is 0.1 times the activation of <flies> (10 * .1 = 1). In simple linear networks, the activation of the unit is simply equal to this value. In more complex networks, the activation of the unit is a nonlinear function of the amount of incoming activation (Rumelhart, Hinton, & Williams, 1986).

Note that the output of this process does not simply specify a single representation. It contains a blend of several partially activated lexical items. Although <BIRD> (the target) is the most strongly activated, its semantic neighbors are partially activated as well. Since activation levels can vary gradiently, connectionist networks allow multiple representations to simultaneously contribute to processing. Note that in this example the cognitive process is structured such that only semantically related items are co-activated; words with nonoverlapping semantic representations (e.g., CAP and DOCK for BIRD) are not activated.

As illustrated above, the structure of a cognitive process – the output generated for a given input – is determined by the spread of activation within the network. In some cases, this is completely specified by the theorist. A cognitive process is hypothesized to have a certain structure; a set of representations, connections, and connection weights is then specified to realize this structure (e.g., Dell, 1986). In other cases, some aspects of the structure of the cognitive process are determined via a learning procedure (e.g., Warker & Dell, 2006).

In networks utilizing a learning procedure, the theorist typically hypothesizes the structure of input and output representations along with the pattern of connectivity within the network (e.g., a set of semantic features fully connected with a set of localist lexical representations). The flow of activation is left unspecified; the network is initialized with a set of random connection weights. The network therefore fails to produce the appropriate output for each input (e.g., given "flying avian pet" it

will activate a set of random lexical items – not "bird" and its semantic neighbors). To allow the network to acquire the correct input–output mapping, the theorist specifies a set of examples of correct mappings between input and output representations (e.g., <flies> <avian> <pet>; <BIRD>). A learning rule uses these data to modify the spread of activation within the network such that after training the network will produce the correct output for each input. As discussed in more detail below, these learning rules typically aim to approximate the covariance between elements of input and output patterns (i.e., the activation over input and output units; van Orden, Pennington, & Stone, 1990). Thus, over the course of learning, connectionist networks acquire the statistical structure of their environment.

Theories of speech production have deployed these general connectionist computational mechanisms to account for a wide variety of empirical patterns in spoken language processing. The following sections examine a few such cases that serve to illustrate the importance of each core connectionist principle.

Gradient Activation in Theories of Speech Production

In connectionist networks, mental representations are realized via gradient patterns of activation over simple processing units. As summarized in Table 7.1, this principle has played a key role in connectionist accounts of an array of empirical phenomena.

Table 7.1 Illustration of Principle 1: mental representations are gradient patterns of activation.

Empirical phenomenon	*Connectionist account*
During phonological spell-out, errors with a mixed semantic-phonological relationship to the target occur at rates greater than chance. Words phonologically related to synonyms of the target show reduced reaction times relative to controls. In bilingual production, cognate words (sharing form and meaning across languages) show reduced reaction times relative to noncognates.	Lexical representations of semantic neighbors (including translation equivalents) are gradiently activated during lexical retrieval; this partial activation cascades to phonological spell-out, facilitating retrieval of their phonological representations.
Phonetic properties of speech errors reflect a trace of the intended target.	Phonological representations of targets remain gradiently activated during production of errors; this partial activation cascades to subsequent phonetic processes, influencing the phonetic realization of the target.

Based on evidence from wide variety of sources, including speech errors (Garrett, 1980), chronometric (reaction time), electrophysiological, and neuroimaging studies (Levelt, 2001), theories of speech production generally assume that there are two major stages of cognitive processing that map meaning (e.g., "flying avian pet") onto the long-term memory representation of its spoken form (the sound sequence /b/ /ɚ/ /d/). The first, lexical selection, is driven primarily by considerations of meaning and grammatical structure. In the context of single word production, this stage involves the selection of a (syntactically appropriate) lexical item to express the intended meaning (e.g., "bird" for "a flying avian pet.") The second stage, phonological spell-out, involves retrieval of the spoken form of this lexical item (e.g., recalling that "bird" contains the sound sequence /b/ /ɚ/ /d/).

Although a large body of evidence suggests these two processing stages form independent components of the cognitive system underlying language production, there is also ample evidence suggesting some degree of interaction between them. One dimension of interaction concerns the activation of semantically related words during phonological spell-out processes. This is supported by the *mixed error effect*. During phonological spell-out processes, errors that have a mixed semantic and phonological relationship to the target (e.g., "shirt" → "skirt") are found to occur more often than predicted based on the rates of purely semantic (e.g., "shirt" → "pants") or purely phonological errors (e.g., "shirt" → "hurt"). This has been documented in speech errors arising spontaneously (Dell & Reich, 1981) or due to acquired impairments to phonological spell-out (Rapp & Goldrick, 2000). This is unexpected if lexical selection and phonological spell-out are completely independent stages of processing. If these two processing stages do not interact, semantic factors should only influence lexical selection – not phonological spell-out. Mixed errors should therefore be no more likely to occur during spell-out than purely phonological errors.

To account for this pattern, theorists have appealed to the connectionist principle of gradient activation (but see Roelofs, 2004). Following the example above (Figure 7.1), assume lexical selection is implemented by a mapping from semantic features to lexical units. Phonological spell-out is implemented by a mapping from lexical to phoneme units. Spreading activation allows multiple semantically related words to become activated during the course of lexical selection. As illustrated above, this causes the output of lexical selection to reflect not just the activation of the target but the partial activation of its semantic neighbors. For example, although the target <SHIRT> will be most strongly activated, the semantic neighbor <SKIRT> will be partially activated (in contrast, the purely phonologically related word <HURT> will be inactive). The spread of activation from the partially activated representations into phonological spell-out processes is referred to as cascading activation. One consequence of this cascade is the activation of the phonological representations of semantically related words. For example, cascading activation from <SKIRT> will activate its constituent sounds. This will boost the activation of the nontarget initial cluster /sk/. In contrast, the nontarget initial consonant /h/ will not receive a similar benefit; "shirt" → "*sk*irt" will therefore

be more likely than "shirt" → "*h*urt." Connectionist simulation studies (Rapp & Goldrick, 2000) show that these mechanisms are sufficient to yield a mixed error effect within phonological spell-out processes.

Cascade from gradiently activated nontarget representations has also been used to account for patterns in chronometric studies (see Goldrick, 2006, for a review). For example, Peterson and Savoy (1998) found faster naming latencies for words phonologically related to the synonyms of target pictures. For example, after presentation of a target picture "couch," facilitation was observed in the subsequent latency for reading aloud words phonologically similar to the synonym "sofa" (e.g., "soda"). Cascading activation provides a ready account of this effect. If the lexical representation <SOFA> is partially activated during processing of the target <COUCH>, cascading activation will provide a boost to its phonological representations (e.g., /s/, /a/). This facilitates processing of words overlapping in form (e.g., "soda").

The cascade of gradient activation of nontarget items has been used to account for a wide range of empirical data. For example, Costa, Caramazza, and Sebastian-Galles (2000) report that bilingual individuals have shorter naming latencies for cognate words (sharing form and meaning across the two languages) relative to noncognates. They attribute this to gradient activation of the lexical representation of the translation equivalent in the nontarget language. Since cognates share the target's phonology, cascading activation from the lexical representation of the translation equivalent facilitates access to the sounds of the target word. For example, for a Spanish–English bilingual the co-activation of <GUITAR> and <GUITARRA> will facilitate access to shared phonological structures (e.g., /g/). In contrast, noncognates provide no support for the target's sounds. For example, <TABLE> and <MESA> do not share sounds; their co-activation will therefore not facilitate retrieval of the target's sounds.

More recent studies have utilized gradient activation to account for interactions between phonological spell-out and subsequent postlexical phonetic processes. Phonetic processes specify the details of how sounds are articulated (e.g., specifying that in English an initial /ba/ is produced by closing the lips, releasing them, and then a short time later allowing the vocal folds to vibrate). As demonstrated by both acoustic (Goldrick & Blumstein, 2006) and articulatory data (McMillan, Corley, & Lickley, 2009), phonetic properties of speech errors reflect properties of both the intended target and the error outcome. For example, when the intended target "big" is replaced by an error "pig" (written as "big" → "pig"), the [p]'s phonetic properties are different than correctly produced instances of "pig." Specifically, in errors resulting in [p] the voice onset time (roughly, the amount of time between the opening of the lips and the beginning of vocal fold vibration) is shorter – more similar to that of [b] – compared to correct productions. This pattern can be attributed to gradient activation of the target's representation during phonological spell-out. The phonological representation of the sound /b/ remains partially active during production of a /b/ → [p] error. This partial activation cascades to postlexical phonetic processing, distorting the articulation of the [p] error to reflect a "trace" of the

intended target's phonetic properties. This account illustrates how gradient activation is a general principle of theories of speech production processing, crossing levels of representation and linguistic populations.

The gradient nature of activation has also allowed researchers to examine not just the simple presence versus absence of interaction but also varying degrees of interaction between cognitive processes. By limiting the relative activation of nontarget representations, it is possible to minimize the degree to which they can influence subsequent processes. This reduces interactive effects, allowing processing to be more discrete. Limiting interaction in this manner allows theories to better account for the full range of speech production data (Goldrick, 2006; Rapp & Goldrick, 2000). For example, chronometric studies have found strong phonological activation only for words that are highly semantically related to the target (e.g., synonyms; Peterson & Savoy, 1998). This pattern can be understood by postulating strong limits on the activation of nontarget items. Only those lexical representations that are strongly activated via spreading activation from semantic representations will be able to influence subsequent processing. Other lexical representations will be too weakly activated to influence phonological spell-out processes – limiting the presence of interactive effects. To implement these limitations on interaction, a number of selection mechanisms have been proposed. They include boosting the target's activation (Dell, 1986) as well as inhibiting nontarget representations (Harley, 1995). Either of these serve to enhance the relative activation of the target versus competing lexical representations.

Spreading Activation in the Speech Production System

The generation of output representations based on input representations is instantiated by the spread of activation between simple processing units. Spreading activation critically interacts with the core representational principles of connectionist networks. Excepting strictly local representations, each processing unit participates in the realization of multiple distinct mental representations. The spread of activation from overlapping representations automatically leads to the activation of a set of related representations. Following the lexical selection example above, overlap in semantic features causes the automatic activation of not just the target "bird" but its semantic neighbors "dog" and "bee" as well. This automatic activation of representations with overlapping representational components has been used to account for a wide array of empirical observations. Table 7.2 provides an overview of accounts utilizing the principles that are reviewed in the section below.

One very general observation regarding errors in speech production is that they tend to result in representations that share structure with the target. For example, errors arising in lexical selection during single word naming tend to share semantic features with the target (e.g., "tiger" → "lion"; Rapp & Goldrick, 2000). Lexical selection errors involving items from the sentence context tend to share grammatical category (e.g., "This *spring* has a *seat* in it"; Garrett, 1980). Similar effects are

Table 7.2 Illustration of Principle 2: cognitive processing as spreading activation.

Empirical phenomenon	*Connectionist account*
Errors tend to result in representations that share structure with the target.	Spreading activation from overlapping processing units automatically boosts the activation of representations that share structure with the target.
During lexical selection, errors with a mixed semantic-phonological relationship to the target occur at rates greater than chance. Phonological errors result in words at rates greater than chance. Phonetic properties of speech errors reflect less of a trace of the intended target when the outcome is a word.	Spreading activation from phonological representations facilitates the activation of lexical representations that share the target's phonology. This enhanced activation cascades to phonological and phonetic levels.
Compared to words that share structure with few lexical items, words that share phonological structure with many lexical items are retrieved more quickly and accurately. Compared to words that share structure with few lexical items, words that share phonological structure with many lexical items are produced with more extreme articulations.	Spreading activation from phonological representations facilitates the activation of lexical representations that share the target's phonology. These lexical representations re-activate the phonological structure of the target, boosting its activation. This enhanced activation cascades throughout the production system.

seen at smaller grain sizes; phoneme-level exchanges tend to involve similar sounds (e.g., "made possible" → "*p*ade *m*ossible"; Shattuck-Hufnagel & Klatt, 1979). Spreading activation offers a natural account of this general phenomenon. Since representations that share structure with the target are automatically activated by spreading activation, they have an inherent advantage over unrelated representations. When processing is disrupted (as in the circumstances that give rise to production errors), this activation advantage means that related forms will be more likely to dominate processing. Following the example of Figure 7.1, during processing of the target "bird," the higher activation of semantically related "bee" means that it is more likely to be mis-selected than the unrelated word "cap."

Goldrick (2008) offers a formal analysis of the conditions under which this account of errors holds. In feature-based or semi-local representational frameworks (such as the semantic features in Figure 7.1), error probability will always reflect representational overlap. Interestingly, in distributed representational frameworks this is not necessarily true. If the distance between the patterns of activation realizing mental representations does not reflect the degree to which their cognitive representations overlap, errors will not respect similarity (see Goldrick, 2008, for

further discussion). For example, if the semantic representations for "bird" and "bee" are realized by activation patterns that are further apart than the patterns for "bird" and "cap," errors may be more likely to produce the unrelated form "cap" than the related form "bee." This highlights the critical interrelationship between spreading activation and representational structure (at both cognitive and connectionist levels of representation) in accounting for the influence of similarity on errors.

Spreading activation has also played a critical role in understanding interactions between stages of processing in speech production. The preceding section discussed how cascading activation from gradiently activated lexical representations allowed semantic relationships to influence phonological processing. Allowing the reciprocal spread of activation – from phonological to lexical representations – allows phonological similarity to influence lexical selection. This mechanism can account for the observation of mixed error effects during lexical selection.

Rapp and Goldrick (2000) report such an effect in the errors of P. W., an individual with an acquired deficit to speech production processes. Although his semantic processing is intact, P. W.'s spoken errors are all semantically related (e.g., naming a picture of a tiger "lion"). This pattern is not predicted by damage to phonological processing; as confirmed by simulation analyses, disruptions to this level would result in purely phonologically related words and nonwords. Rapp and Goldrick therefore attribute P. W.'s errors to a disruption of lexical selection processes. Critically, P. W.'s semantic errors have a higher degree of phonological overlap than would be predicted by random substitutions of words within a semantic category. This suggests that errors sharing both phonological and semantic structure – mixed errors – are more likely to occur than purely semantically related errors.

This pattern can be accounted for by allowing activation to spread from phonological to lexical representations. This feedback from "late" to "early" processing stages boosts the activation of mixed errors during lexical selection. For example, the target "bird" activates the phoneme /b/. Feedback from this boosts the activation of the mixed semantically and phonologically related word "bee," producing an activation advantage for it over the purely semantically related word "dog." This activation advantage leads to a mixed error effect following disruptions to lexical selection (Rapp & Goldrick, 2000).

Spreading activation from phonological representations also provides a mechanism to account for the influence of purely phonologically related lexical representations (e.g., for target "mitten," <MUFFIN>). These are frequently referred to as *neighbors* of the target. In the absence of feedback, such words would not become active; semantically driven activation from the speaker's intended message would only activate words that share the target's meaning. Feedback from shared phonological structure is required to activate the lexical representations of these purely phonologically related neighbors.

Cascade from the lexical representations of neighbors enhances the activation of their phonological representations. This provides an advantage for phonological representations corresponding to words over those corresponding to nonwords. For example, for target "bird" the lexical representation of the neighbor <HERD> will

be strongly activated by feedback by the shared phonemes /ɚ/ and /d/. The lexical representation <HERD> will then enhance the activation of /h/. In contrast, the phonological representation of the nonword outcome "zerd" will receive little support; there is no lexical representation *<ZERD> to provide strong activation to the phoneme /z/. This accounts for the *lexical bias effect*, the observation that phonological errors are more likely to result in words than predicted by chance (Dell, 1986; but see Levelt, Roelofs, & Meyer, 1999).

This strengthening of word error outcomes persists into subsequent phonetic processes. As discussed above, phonetic properties of speech errors reflect properties of both the intended target and the error outcome (e.g., the [p] in the error "big" → "pig" exhibits a phonetic trace of the intended target /b/). The influence of the intended target is significantly smaller for word versus nonword error outcomes. For example, the [p] in "bid" → "pid" will exhibit a greater trace of the target /b/ than the [p] in "big" → "pig" (Goldrick & Blumstein, 2006; McMillan et al., 2009).

Spreading activation from phonological representations serves not only to strengthen nontarget representations; it also facilitates access to the target. Specifically, words that are similar to many lexical items show a processing advantage relative to those that are similar to few words. Words with many neighbors are produced with shorter latencies than words with few neighbors (Vitevitch, 2002). Following acquired deficits affecting phonological spell-out, words with few neighbors are produced less accurately than words with many neighbors (Goldrick, Folk, & Rapp, 2010). Additionally, words with many neighbors are less likely to fall into tip-of-the-tongue (TOT) states (Harley & Bown, 1998). The enhanced activation of words with many neighbors manifests itself not just in speed and accuracy but also in articulation. Relative to words with few neighbors, words with many neighbors are produced with more extreme articulations (consonants: Baese-Berk & Goldrick, 2009; vowels: Wright, 2004).

Feedback-driven activation of lexical representations provides an account of the facilitatory influence of neighbors. After the lexical representations of neighbors are activated by feedback, they in turn reactivate the phonological structure they share with the target. Feedback allows this enhanced activation to strengthen the lexical representation of the target (Dell & Gordon, 2003). Cascade then serves to strengthen processing of the target at phonological and phonetic processing levels. For example, target <BIRD> activates <HERD> via feedback from the shared phonemes /ɚ/ and /d/. <HERD> in turn reactivates /ɚ/ and /d/; via feedback, these provide an additional boost to the target <BIRD>. Via cascade, the heightened activation of <BIRD> influences processing of its phonological and phonetic structure – enhancing the speed and accuracy of retrieval and leading to more extreme articulations.

Acquiring the Statistical Structure of the Environment

The preceding discussion has shown how spreading activation is a central principle of connectionist explanation. The final principle examined here focuses on what

Table 7.3 Illustration of Principle 3: cognitive processes are structured via statistical learning.

Empirical phenomenon	*Connectionist account*
Speech errors tend to not result in sequences of sounds that are infrequent within an experimental context.	Spreading activation is continuously updated to reflect the statistical structure of sound sequences. Infrequent sound sequences are therefore less activated than frequent sequences.
Picture naming is slower and less accurate following the retrieval of semantically related items.	Spreading activation is continuously updated to reflect the covariance of semantic features and lexical representations. Recently produced lexical items are therefore more strongly reactivated by their semantic features.

determines the flow of activation. In the majority of theories reviewed above, the spread of activation is stipulated by the theorist. The theorist specifies one or more hypotheses regarding the structure of some cognitive process and then implements this hypothesis using spreading activation. In contrast, the research reviewed below uses statistical learning algorithms to address the origin of this structure. Table 7.3 provides an overview of these contributions.

Warker and Dell's (2006) proposal provides an illustration of this approach. They focus on the linking of segmental units or phonemes (retrieved during phonological spell-out) with prosodic structure (specifically syllables). On the basis of speech error and chronometric data, most current theories of speech production (e.g., Levelt et al., 1999) assume that such information is not linked together in long-term memory. The output of phonological spell-out specifies the linear order of a set of segments; these must be associated with prosodic structure before they can engage articulatory planning and execution processes. Warker and Dell therefore postulate a syllabification process that takes as an input a set of phonemes and a specification of their linear order; it generates as output an association of segments to syllable positions (e.g., /b/-onset /ɚ/-nucleus /d/-coda for "bird").

To realize this process as a connectionist network, Warker and Dell (2006) construct two semi-local levels of representation. The input representation has a single unit for each segment, followed by a unit representing the linear order of those segments. For example, for "cat" the representation would be: /æ/ /k/ /t/ <k-1 æ-2 t-3>. For "zat" it would be /æ/ /t/ /z/ <z-1 æ-2 t-3>. The output representation has a single unit for each segment within each syllable position. For example, the "cat" representation would be </k/-ONSET> </æ/-VOWEL> </t/-CODA>.

Rather than specify how activation spreads within the network, Warker and Dell (2006) allowed the network to acquire structure through learning. Initially the connections are seeded with random values. The network is then presented with a series of correct input–output representation pairs. The input is presented and the network generates an output. This is compared to the correct output. If the network's

response is different than the target output, the weights throughout the network are slightly adjusted (using the algorithm in Rumelhart et al., 1986) such that the next time the input is presented the network is more likely to generate the correct output. Critically, the algorithm learns to perform this mapping by encoding the statistical structure of the training set – specifically, the covariance between the activation of input and output units (van Orden et al., 1990).

Warker and Dell's (2006) network also incorporates a set of *hidden* units that intervene between the input and output representations. These are hidden in the sense that they do not interface with processing "external" to the syllabification process; furthermore, their activation values are not specified by the theorist. The activation patterns over these units emerge as a consequence of the learning algorithm that adjusts the spread of activation. Incorporating such units greatly increases the range of input–output mappings that connectionist networks can compute (Rumelhart et al., 1986), making them a standard feature of connectionist theories incorporating learning.

After the learning rule has adjusted the weights for several thousand input–output pairs (reflecting the full range of English monosyllables), Warker and Dell's (2006) network assigns a strong activation to the correct segment–syllable position bindings and low activation to the incorrect outputs. For example, for syllables like "cat" it assigns 0.9 to /k/-onset but only 0.05 to /k/-coda. This shows that this network can acquire the correct structure to compute the syllabification process. Of greater interest, however, is the network's ability to encode the statistical regularities that hold across this set of syllables. For example, when presented with an /h/ as input, the network should strongly activate /h/-onset but not /h/-coda. The latter never appears in any English syllables (e.g., there are English words like "him" but none like *"mih"). In fact, Warker and Dell report that for target "hing," the error /h/-coda has an activation of only 0.002. In contrast, for target "kaf," the error /k/-coda has an activation of 0.05 (much closer to the 0.9 activation level of the correct segment–syllable position bindings). The weaker activation of /h/-coda reflects the network's encoding of the statistical structure of English.

This learning mechanism provides an account for the tendency of speech errors to respect phonotactic regularities (statistical regularities in sound sequences). As noted above, in English no syllables end in /h/. When English speakers make speech errors, the errors rarely result in the production of /h/ in coda position (i.e., very few errors resemble "head lock" → "head lo*h*"; Vousden et al., 2000). Assuming that relative activation reflects production probability, Warker and Dell's (2006) proposal provides an account of how syllabification processes come to respect these phonotactic regularities (see Goldrick, 2007, for a more formal analysis of this proposal). Because spreading activation reflects the statistical structure of the environment, the errors produced during processing will be sensitive to the distributional properties of sound sequences.

Warker and Dell (2006) propose that this learning mechanism can also account for speakers' sensitivity to shifts in the statistical structure of their environment. Dell, Reed, Adams, and Meyer (2000) document such a shift by showing that par-

ticipants can acquire novel phonotactic regularities. In Dell et al.'s study, participants read aloud sets of tongue twisters that exhibited novel phonotactic regularities. For example, in contrast to the participant's native language (English), /f/ appeared only in the onset of syllables, never in the coda (e.g., syllables like "feg" were presented, but "geff" was not). Other segments respected their normal distribution in English, appearing in both onset and coda position. The participant's speech errors reflected these novel regularities. When a segment was restricted to one position, errors overwhelmingly favored producing the segment in that position. For example, when /f/ was restricted to the onset of syllables, onset errors such as "feg kem" → "feg *f*ek" made up 97% of errors resulting in /f/. In contrast, for segments that appeared in both onset and coda, errors appeared in both syllable positions. For example, /m/ appeared in both onset and coda; errors such as "meg kef" → "meg *m*ef" made up approximately 70% of errors involving this segment.

Warker and Dell (2006) account for the acquisition of these novel regularities by appealing to the continual operation of the learning mechanism outlined above. To model the experience of the experiment, following exposure to all of the syllables of English the connectionist network received additional training on the experimental syllables. The network was therefore exposed to many syllables in which /f/ was associated to onset, but no syllables where /f/ was associated to coda. Similar to the phonotactic regularities observed in the training set as a whole, spreading activation within the network changed to reflect these new statistical regularities. For example, if the target was "feg," the error /f/-coda had an activation of 0.02. In contrast, following this additional training the error /g/-coda (for input "gef") had an activation of 0.04. Through continual operation of a statistical learning procedure, network processing has adapted to reflect the statistical structure of the environment – providing an account for the flexible nature of syllabification processes in speech production.

Furthermore, Warker and Dell's (2006) connectionist learning mechanism provides an account of the relative difficulty of acquiring different statistical regularities. Warker and Dell report behavioral findings that show more complex regularities (e.g., /f/ is onset when the vowel is /i/, but is the coda when the vowel is /e/) require more time for participants to learn than simple segment–syllable position regularities (e.g., /f/ is always in onset). The connectionist learning mechanism exhibits this same pattern; it requires greater training to acquire the more complex phonotactic regularities (see Goldrick, 2007, for further discussion).

Dell, Oppenheim, and Kittredge (2008) adopt a similar learning-based perspective to account for the dynamics of lexical selection. Behavioral studies have shown that the difficulty of lexical selection can vary depending on the relationship between items that have been recently processed. When participants are asked to name a series of pictures from a single semantic category, they are slower and less accurate compared to trials where the same pictures are named in mixed sets. For example, naming a picture of a fox in the context of a tiger, dog, and owl is slower and less accurate than naming the same picture in the context of a chair, star, and hammer. Because this interference effect is not observed in reading aloud of the same items

(a task which does not require access to semantic representations), this effect has been attributed to semantically driven lexical selection processes (Damian, Vigliocco, & Levelt, 2001). Specifically, context is assumed to enhance competition from semantically related items. As discussed above, spreading activation automatically activates semantic neighbors of the target during lexical selection. If these already-active words are further boosted by the context, they will induce significant competition during lexical selection.

How does context enhance the activation of lexical items? One account is based around priming; prior access to a lexical representation temporarily increases its activation. Such an account has difficulty explaining why the strength of interference is not influenced by the temporal lag between presentation of semantically related items (Howard, Nickels, Coltheart, & Cole-Virtue, 2006). A temporary, activation-based mechanism would be expected to show decay over time.

Dell et al. (2008) offer a learning-based alternative in the same spirit as Warker and Dell (2006). They assume that spreading activation between semantic features and lexical units is driven by a covariant learning rule; furthermore, this learning rule continues to update spreading activation during adult processing. This allows spreading activation to be influenced by the statistical structure of recent trials. Prior presentation of a lexical item strengthens the connections between the corresponding semantic features and lexical unit. This leads the lexical unit to become more active when another item sharing its features is presented – inducing increased semantic interference. Furthermore, because this mechanism changes the structure of lexical selection processes (modifying the flow of activation), it is not subject to a purely temporal decay process; it persists until a sufficiently large number of intervening trials have altered the statistical structure of the input–output mapping. To demonstrate the adequacy of this account, Dell et al. (2008) present a series of simulations showing this approach provides a qualitative match to both latency and error data from this experimental paradigm.

In sum, connectionist theories assume that the structure of cognitive processes – the way in which activation spreads between simple processing units – reflects the statistical structure of the environment. The continuous operation of learning algorithms that allow networks to encode this statistical structure allows connectionist researchers to account for the dynamic, context-dependent nature of speech production processes.

Learning and the Emergence of Structure

Acquiring the structure of cognitive processes via learning algorithms means that connectionist networks are not simply devices for instantiating cognitive theories. They also offer the possibility of developing novel theoretical accounts – acquiring internal structure that differs in substantive ways from existing proposals. To illustrate this approach, the discussion below focuses on the work of Dell, Juliano, and Govindjee (1993), which aimed to develop a new framework for phonological spell-

out processes via a statistical learning with simple recurrent networks (Elman, 1990; Jordan, 1986).

Dell et al. (1993) began with a set of slightly different representational assumptions than the processing architecture discussed in previous sections. First, they assumed that input representations consisted of a distributed pattern of activation encoding lexical identity (rather than strictly local lexical representations). Instead of phoneme units, they assumed the output was a set of phonological features. Finally, and most critically, rather than generating all phonological elements of a word in parallel, the network was trained to sequentially generate the features of each segment. The simple recurrent network architecture provides an appropriate connectionist mechanism for this task. Specifically, in Dell et al.'s networks, the activation of hidden and output units was influenced by the activation of these units at the previous point in time – allowing for recurrent activation flow. Allowing previous states of the network to influence processing endows the network with a "memory" of what occurred previously; this gives simple recurrent networks the capacity to produce sequences of output representations.

The aim of Dell et al.'s (1993) simulations was to develop an alternative to existing accounts of phonological spell-out. These existing accounts make use of frame representations – explicit representations of structure that serve to guide the selection of phonological representations. For example, Dell (1986) assumes a syllabic frame representation consisting of a set of abstract consonant and vowel units. These guide selection of phonemes in the appropriate sequence to generate syllables (e.g., ensuring that for target "bird" the first segment selected is an onset consonant). Critically, these representations form a set of mechanisms distinct from spreading activation from lexical to phoneme units. The structure of the syllable (its consonant/vowel organization) is therefore processed separately from its content (the specific segments that occur in each structural position). (Note, however, that structure processing occurs in parallel with retrieval of content; see below for further discussion.) In contrast to this frame-based model, Dell et al.'s (1993) simple recurrent network utilizes a single set of hidden and context units to both activate the appropriate phonological representations and produce them in the correct sequence. A single set of processing mechanisms therefore implements both structure and content processing.

Dell et al. (1993) find that this network is able to capture empirical patterns that have been attributed to the influence of the syllabic frame. For example, speech errors tend to respect syllabic constituency. Theories of English syllable structure (Venneman, 1988) have postulated that syllables are composed of two constituents, the onset (roughly, all segments preceding the vowel) and the rime (roughly, the vowel and all following segments). Speech errors respect this division; errors involving the replacement of the vowel and coda (e.g., "read" → "rope") are more frequent than errors involving the replacement of the vowel and onset (e.g., "read" → "load"; Stemberger, 1983). This has typically been attributed to the influence of an explicit structural frame that includes a representation of onset–rime structure (Sevald, Dell, & Cole, 1995). However, Dell et al.'s simulations exhibited the same pattern;

the authors take this to suggest that phonological spell-out processes do not need to incorporate explicit structural representations.

Some theorists have questioned whether this account can capture the full range of speech production data (Sevald et al., 1995). Assuming that the proposal is in fact empirically adequate, the challenge is to precisely articulate the content of the connectionist proposal. What is the nature of the network's solution to the problem? Until this has been established to a sufficient level of detail, it is unclear what specific elements of previous theoretical approaches are challenged by these findings (McCloskey, 1991). The complexity of connectionist networks – the distribution of computation over many units and connections – has made such analyses difficult. However, simple recurrent networks have been extensively analyzed in the context of sentence processing using the prediction task (for recent reviews, see Elman, 2009; Tabor, 2009). The task of these networks is to generate predictions regarding the next word in a sentence, given the preceding word (e.g., given the English sequence "The dog bites the . . ." the network should predict a noun will appear). In many cases, these networks converge on a solution where the structural position of elements is encoded by regions within the space of hidden unit activations. The dynamics of the network – how it transitions from one region of the hidden unit space to another – encode the relations between these structural positions (e.g., the serial order or dependencies such as subject/verb number agreement). The regions that define structural positions are themselves structured to encode the particular element that occurs within this position. These analyses suggest that it is possible that Dell et al.'s (1993) networks developed an internal structure that reflected the distinction between structure and content. Structure reflects the high-level dynamical organization of the network's trajectory through hidden unit space; content is reflected by its fine-grained structure. At one level, this agrees with a central property of frame-based models; namely, that structure and content are distinctly represented within the production system. In fact, it is possible that Dell et al.'s networks converged upon a dynamical organization that closely approximated the frame-based proposals – where content and structure are processed in parallel yet independent processing streams.

Research in sentence production has underscored the fact that connectionist networks must acquire sufficient distinctions between structure and content to account for the full range of empirical data. Chang (2002) compared several sentence production architectures built around simple recurrent networks. The architecture analogous to that of Dell et al.'s (1993) phonological spell-out model mapped from a static representation of the speaker's message to a sequence of words. Among other issues, this model failed to exhibit sufficient (and appropriate) generalization from its training set. For example, it had difficulty producing combinations of adjectives and nouns that had not occurred in its training set (e.g., although it could produce *silly dog* and *good cake*, it could not produce *silly cake* – a nonsensical but syntactically legitimate string). In order to acquire an appropriate internal structure that supported generalization, Chang adopted representational and architectural assumptions that built on previous frame-based production theories. First, he

adopted compositional semantic representations; these distinguished event roles (e.g., <agent>) from the lexical concepts that filled those roles (e.g., <man>). In essence, this creates an abstract semantic frame. The second set of assumptions was architectural. Chang assumed that two interacting pathways were involved in the generation of the next word in a sentence. A meaning system mapped directly from the conceptual representation of the sentence. A second sequencing system was a simple recurrent network; it made use of additional sets of hidden units to develop representations that abstract over word categories (much like a syntactic frame). These assumptions enabled the network to exhibit proper generalization and account for a wide range of data from speech production (see also Chang, Dell, & Bock, 2006, for discussion). Critically, analysis of the network's internal structure (Chang, 2002) reveals that this success is made possible in part by developing representations and mechanisms that make robust distinctions between structure (i.e., syntactic role) and content (i.e., lexical items).

Although these architectures do incorporate key insights of more traditional frame-based models, at finer-grained levels of analysis they may be capable of accounting for empirical data that cannot be modeled by traditional frame-based theories. For example, research in sentence processing (Elman, 2009) and in routine sequential actions more generally (Botvinick & Plaut, 2004) have examined how simple recurrent networks exhibit greater expressive power than models built around simple abstract frames. Simple recurrent networks can also capture "quasi-hierarchical" patterns of performance, wherein dependencies between items in a sequence are not limited to relations between categories. For example, Lee and Goldrick (2008) found that English speakers are sensitive not only to onset–rime structure (holding over all onset consonants) but also to relationships between *particular* onset consonants and vowels. In a strict consonant–vowel frame-based framework (as in Dell, 1986), such dependencies could not be represented; the frame collapses all onset consonants into a single entity.

Utilizing connectionist networks and learning algorithms to develop novel theoretical accounts such as those based around quasi-hierarchical structure is very much an open area of speech production research. As shown by the discussion above, detailed analysis of connectionist processing is required. Although they can acquire internal structure that differs in substantive ways from existing proposals, connectionist networks can only contribute to theory development if their internal structure and processing principles are fully explicated.

Conclusion

Connectionist principles have played an integral role in the development of theories of speech production. From semantically driven lexical selection processes to phonetic processes, researchers have accounted for empirical patterns by appealing to connectionist computational mechanisms. In many ways, these mechanisms represent a theoretical vocabulary more powerful than purely symbolic computational

mechanisms. Rather than being limited to digital symbolic representations, connectionist networks can express gradient blends of multiple representational states. Furthermore, as discussed in the preceding sections, connectionist networks offer a procedure for developing internal representations; in contrast, many symbolic architectures are forced to postulate representational structure a priori.

An important challenge for connectionist theories is to rein in this expressive power. This chapter touched on two such points. Gradient activation flow and the spread of activation must be restricted; empirical evidence suggests there are strong limitations on the strength of interactions between stages of spoken production (Rapp & Goldrick, 2000). Although digital representations are too restrictive, they may in fact represent a close approximation of the state of many stages of processing. The second issue reviewed above suggests a similar conclusion. The work of Chang and colleagues (Chang, 2002; Chang et al., 2006) underscores the need for abstract, symbol-like representations to account for many of our speech production abilities. Again, symbolic representations may provide a good approximation for the structure of cognitive processes. The connectionist challenge is to explain the constraints on learning, processing, and behavior that give rise to this structure.

Note

Thanks to Melissa Baese-Berk for helpful comments. Preparation of this chapter was supported by NIH grant NIDCD DC007977 and NSF grant BCS-0846147.

References

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, *24*, 527–554.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395–429.

Chang, F. (2002) Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609–51.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1283–1296.

Damian, M. F., Vigliocco, G., & Levelt, W. J. M. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*, B77–B86.

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities*. New York: Mouton de Gruyter.

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*, 149–95.

Dell, G. S., Oppenheim, G. M., & Kittredge, A. K. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and Cognitive Processes*, *23*, 583–608.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1355–1367.

Dell, G. S. & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*, 611–629.

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

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 547–582.

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production: Speech and talk* (vol. 1, pp. 177–220). New York: Academic Press.

Goldrick, M. (2006). Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence. *Language and Cognitive Processes*, *21*, 817–855.

Goldrick, M. (2007). Constraint interaction: A lingua franca for stochastic theories of language. In C. T. Schütze & V. S. Ferreira (Eds.), *The state of the art in speech error research: Proceedings of the LSA Institute workshop* (MITWPL vol. 53, pp. 95–114). Cambridge, MA: MIT Working Papers in Linguistics.

Goldrick, M. (2008). Does like attract like? Exploring the relationship between errors and representational structure in connectionist networks. *Cognitive Neuropsychology*, *25*, 287–313.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*, 649–683.

Goldrick, M., Folk, J., & Rapp, B. (2010). Mrs Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*, 113–134.

Harley, T. A., (1995). Connectionist models of anomia: A comment on Nickels. *Language and Cognitive Processes*, *10*, 47–58.

Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, *89*, 151–174.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*, 464–482.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach*. Institute for Cognitive Science Report 8604. University of California, San Diego. Reprinted (1997) in J. W. Donahoe & V. P. Dorsel (Eds.), *Neural-network models of cognition: Biobehavioral foundations* (pp. 221–277). Amsterdam: Elsevier.

Lee, Y., & Goldrick, M. (2008). The emergence of sub-syllabic representations. *Journal of Memory and Language*, *59*, 155–168.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*, 13464–13471.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, *2*, 387–395.

McMillan, C., Corley, M., & Lickley, R. (2009). Articulatory evidence for feedback and competition in speech production. *Language and Cognitive Processes*, *24*, 44–66.

Palmer, S. E., & Kimchi, E. (1986). The information processing approach to cognition. In T. J. Knapp & L. C. Roberts (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 37–77). Hillsdale, NJ: Erlbaum.

Peterson, R. R., & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 539–557.

Rapp, B. & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*, 460–499.

Roelofs, A. (2004). Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). *Psychological Review*, *111*, 561–572.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group *Parallel distributed processing: Explorations in the microstructure of cognition* (vol. 1, *Foundations*, pp. 318–62). Cambridge, MA: MIT Press.

Sevald, C. A., Dell, G. S., & Cole, J. (1995). Syllable structure in speech production: Are syllables chunks or schemas? *Journal of Memory and Language*, *34*, 807–820.

Shattuck-Hufnagel, S., & Klatt, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech errors. *Journal of Verbal Learning and Verbal Behavior*, *18*, 41–55.

Smolensky, P. (1999). Grammar-based connectionist approaches to language. *Cognitive Science*, *23*, 589–613.

Smolensky, P. (2006). Computational levels and integrated connectionist/symbolic explanation. In P. Smolensky & G. Legendre *The harmonic mind: From neural computation to optimality-theoretic grammar* (vol. 2, *Linguistic and philosophical implications*, pp. 503–592). Cambridge, MA: MIT Press.

Stemberger, J. (1983). *Speech errors and theoretical phonology: A review*. Unpublished manuscript, Carnegie-Mellon University. Distributed by the Indiana University Linguistics Club, Bloomington, IN.

Tabor, W. (2009). Dynamical insight into structure in connectionist models. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), *Towards a unified theory of development: Connectionism and dynamical systems theory reconsidered* (pp. 165–181). Oxford: Oxford University Press.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, *97*, 488–522.

Venneman, T. (1988). *Preference laws for syllable structure and the explanation of sound change*. Berlin: Mouton de Gruyter.

Vitevitch , M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 735–747.

Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, *41*, 101–75.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 387–398.

Wright, R. A. (2004). Factors of lexical competition in vowel articulation. In J. J. Local, R. Ogden, & R. Temple (Eds.), *Laboratory Phonology VI* (pp. 26–50). Cambridge: Cambridge University Press.