Modeling the relation between the production and recognition of spoken word forms

Ardi Roelofs

> In de spraak vinden het gemoeds- en het verstandsleven, beide, den klaarsten vorm van uitdrukking, die wederkeerig op de ontwikkeling van beide krachtig terugwerkt [In speech, our emotional and intellectual lives find their clearest form of expression, which in turn forcefully feeds back on the development of both] (Donders 1870: 10)

1. Introduction

The production of spoken words and their recognition have been intensively investigated in psycholinguistics during the past several decades. On the one hand, spoken word recognition has been investigated using tasks such as cross-modal semantic priming (e.g., Swinney et al. 1979; Warren 1972), auditory lexical decision (e.g., McCusker, Hillinger, and Bias 1981), speech shadowing (e.g., Cherry 1957; Marslen-Wilson 1973), phoneme monitoring (e.g., Foss 1969; Frauenfelder and Segui 1989), gating (e.g., Grosjean 1980), and word spotting (e.g., Cutler and Norris 1988). Furthermore, research has used eye-tracking techniques to monitor participants' eye-movements as they follow spoken instructions to manipulate real objects (e.g., Tanenhaus et al. 1995). Spoken word recognition has also been studied using neuroimaging techniques (see Price, Indefrey, and Van Turennout 1999, and Hickok and Poeppel 2000, for reviews).

On the other hand, spoken word production has been investigated mainly through the analyses of corpora of naturally occurring speech errors (e.g., Dell and Reich 1981; Fromkin 1971; Garrett 1975; Shattuck-Hufnagel 1979; Stemberger 1985) and, more recently, in experiments using the picture naming task and Stroop-like paradigms such as picture-word interference (e.g., Glaser and Düngelhoff 1984;

Jescheniak and Levelt 1994; Lupker 1979; Roelofs 1992; Schriefers, Meyer, and Levelt 1990; Stroop 1935; see Levelt 1989, and MacLeod 1991, for reviews). The measurement of interest is usually the time it takes to name the pictures, although studies have also examined the naming errors that are occasionally made (e.g., Martin et al. 1996), the eye-gaze durations using a head-mounted eye-camera (e.g., Griffin and Bock 2000; Meyer et al. 1998; Meyer and Van der Meulen 2000), and the brain areas involved (e.g., Levelt et al. 1998; see Indefrey and Levelt 2000, for a meta-analysis of 58 neuroimaging studies of word production).

The empirical investigations have lead to the development of detailed computationally implemented models of spoken word recognition (see Norris, McQueen, and Cutler 2000a, 2000b for discussion), and spoken word production (see Levelt 1989, 1999 for reviews). However, the relationship between production and recognition has received surprisingly little attention (see Monsell 1987, for a review). Yet, an examination of the literature in the recognition and production domains reveals that both lines of research distinguish between levels of phonological features, phonemes, and words in form-based processing. Furthermore, the cognitive neuroscience literature, focussing independently either on speech recognition or production, has identified a brain area, the left posterior superior temporal lobe, that participates in the phonemic level of processing in both speech perception and production (see Buchsbaum, Hickok, and Humphries 2001, for a review). This raises the question whether a single system participates in phonetic and phonological processing during both production and recognition (e.g., Allport 1984; MacKay 1987) or whether there are separate phonetic and phonological systems for production and recognition (e.g., Dell et al. 1997; Levelt, Roelofs, and Meyer 1999a). This issue is addressed in the current chapter.

Although the computational models independently developed for speech production and recognition have addressed several types of data sets, most models (with as only exception the error-based models of spoken word production) have attempted to account for chronometric findings, such as speech production and recognition latencies, eye-gaze durations and the distribution of eye-fixation

probabilities over time, and the time course of brain activation. Interestingly, the first person to measure speech recognition and production latencies, Donders (1868), also developed a model for eye movements and examined cerebral blood flow, which, together with a subtractive method he designed, underlies two of the most widely used modern functional neuroimaging techniques in speech recognition and production research, PET (positron emission tomography) and fMRI (functional magnetic resonance imaging). Donders was also interested in the mechanisms underlying speech. In his monograph "De physiologie der spraakklanken, in het bijzonder van die der Nederlandsche taal" [The physiology of speech sounds, in particular those of the Dutch language] (Donders 1870), he gave a detailed account of the acoustic and phonetic properties of (Dutch) speech sounds and how they are articulated. In his chronometric work, Donders held that mentally progressing from hearing speech to producing speech involves a *translation* process, that is, the mental processes dealing with speech input and output are different. Donders lacked, however, the theoretical apparatus to precisely specify and develop his ideas about mental processes — the basics of the computational theory of mind (and its modeling tools) he would have needed took the full first half of the twentieth century to develop.

At the time Donders conducted his revolutionary chronometric studies, Wernicke (1874) made the seminal observation that brain-damaged patients with speech recognition deficits (today called Wernicke's aphasics) often have fluent but phonemically disordered speech production. Based on a post-mortem examination of one of his patient's lesion site, Wernicke proposed that the left posterior superior temporal lobe of the human brain stores "auditory word images" that are activated in both speech recognition and production, and that these auditory images are translated into "motor word images" (presumed to be stored in frontal areas) during speech production. The activation of the auditory word images during speech production was supposed to assist the selection of the appropriate motor word images. Consequently, when the auditory word images are lesioned (as in Wernicke's aphasia), or when the anatomical pathways connecting auditory and motor systems are disrupted (as Wernicke

assumed in the case of conduction aphasia), the selection of motor images was assumed to be no longer appropriately constrained, explaining the phonemic paraphasias.

Almost a century later, at the end of the 1960s, Liberman et al. (1967) proposed a motor theory of speech perception, which holds that the target representations of speech perception are the very same articulatory motor programs that are used for speech production. No translation is necessary to go from perception to production. And two decades later, MacKay (1987) developed a general theory in which spoken language comprehension and language production are accomplished in their entirety by one and the same system.

In this chapter, I address the issue of shared versus separate systems for speech recognition and production within the context of computationally implemented models of spoken word recognition and production, specifically TRACE (McClelland and Elman 1986), Shortlist (Norris 1994), the DSMSG model (Dell 1986; Dell et al. 1997), and WEAVER++ (Levelt et al. 1999a; Roelofs 1992, 1997a). Due to space limitations, other models such as the unimplemented model of Caramazza (1997) are not discussed. A problem with MacKay's (1987) theory for present purposes is that it is rather speculative and that it has not been specified computationally, which makes it difficult to evaluate the implications. The claims of Liberman et al. (1967) mainly concerned early aspects of speech perception, whereas I focus on spoken word recognition and production in this chapter. In particular, I concentrate on the representation and processing of word forms. Issues concerning word forms are to a certain extent independent of higher-order aspects of speech. For example, Levelt et al. (1999a) argued for separate form-based systems for speech recognition and production, but for shared syntactic and semantic systems.

In what follows, I make a case for a modern version of Donders' original position of closely linked but distinct mental systems for speech recognition and production as far as word forms is concerned. After a short excursion to Donders' pioneering work measuring the latencies of speech production and recognition, I discuss some of the most important computationally implemented models of spoken word recognition and production, with an eye on their time course

characteristics and the relation between recognition and production, along with some key empirical findings supporting the models. In particular, I briefly discuss the most prominent recognition model that assumes feedback (TRACE) and the model that does not (Shortlist), and the most prominent production model that assumes feedback (DSMSG) and a model that does not (WEAVER++). All four models achieve form processing through activation networks.

For spoken word recognition and production to be subserved by the same system of representations and processes, the presence of bottom-up phoneme-to-word links (for recognition) and top-down word-to-phoneme links (for production) in the system is a necessary condition. The existence of top-down links in a spreading activation network for production implies activation feedback in the same network during recognition, and the existence of bottom-up links for recognition implies feedback during production. If there is no good evidence for feedback in both recognition and production, then it is unlikely that recognition and production are achieved by the very same system.

Feedback is not a sufficient condition for a shared system, however. Form recognition and production may be achieved by separate systems, each including feedback (cf. DSMSG). Furthermore, although in interactive models like TRACE and DSMSG, feedback occurs mandatorily, there is the logical possibility that in a shared recognition/production system, the bottom-up links may be operative only during actual word recognition and the top-down links only during actual word production (cf. Norris et al. 2000b). To evaluate this latter possibility, evidence from combined recognition/production tasks rather than from pure recognition or pure production tasks is critical. Three such tasks are auditory picture-word interference, auditory lexical decision during object naming, and auditory priming during speech preparation. Evidence from these tasks is discussed. Finally, I discuss evidence from recent functional neuroimaging studies examining Wernicke's claim that exactly the same brain area, the posterior superior temporal lobe, participates in both speech recognition and production. Alternatively, different subregions of this broad area could be involved. I conclude that the available evidence

supports the idea of separate but closely linked feedforward-only systems for word-form production and recognition.

2. Donders' ground-breaking work

Donders' work in the nineteenth century has in many respects anticipated the modern experimental study of speech production and recognition. His techniques and views are strikingly modern. Donders took great interest in eye movements, for which he developed a mechanical model. Furthermore, he investigated cerebral circulation (e.g., Donders 1849) and highly valued the discovery of the metabolism of the brain suggesting its action. “As in all organs, the blood undergoes a change as a consequence of the nourishment of the brain”. One “discovers in comparing the incoming and outflowing blood that oxygen has been consumed” (Donders [1868] 1969: 412). This latter insight, together with a task-subtractive method designed by Donders, underlies the two most widely used functional imaging techniques, PET and fMRI. Donders published on natural selection in 1848, some ten years before Darwin's “The origin of species” appeared in print. Donders realized that the mind is not the brain, but is what the brain does: “A complete knowledge of the functioning of the brain, with which each mental process is connected, does not carry us a step further in the understanding of the nature of their relation” (Donders [1868] 1969: 412). He lacked the formal language to specify mental processes precisely, but discovered another handle on them: response time. Until then, the received view held that the mental operations involved in responding to a stimulus occur instantaneously. “But will all quantitative treatment of mental processes be out of the question then? By no means! An important factor seemed to be susceptible to measurement: I refer to the time required for simple mental processes” (Donders [1868] 1969: 413-414).

Donders attempted to describe the processes going on in the mind by analyzing cognitive activity into separate, discrete stages that the brain goes through when faced with different tasks. To this end, he measured, among other things, speech production latencies and had

participants respond to spoken stimuli by manual key-press responses.

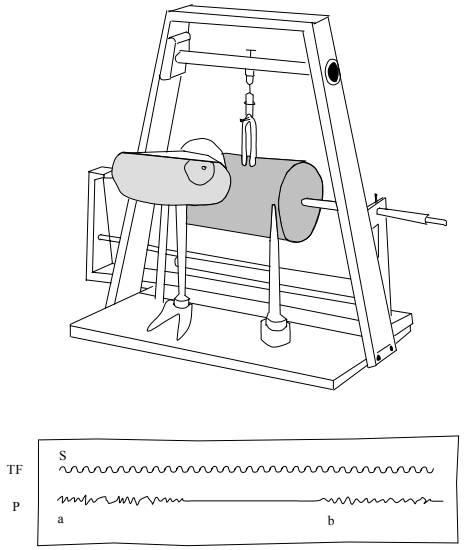

Figure 1. The "noematachograph and phonautograph" designed by Donders to measure speech recognition and production latencies. See the text for an explanation of their working.

Figure 1 provides a sketch of the "noematachograph and phonautograph" designed by Donders to measure speech recognition and production latencies. An experimental trial proceeded as follows. Two participants A and B were seated before the mouth of the phonautograph. While the cylinder was rotated, A uttered a syllable and B had to repeat it as quickly as possible without making mistakes (cf. Cherry 1957; Marslen-Wilson 1973). The beginning of the oscillations caused by each of the two sounds was marked on paper by points *a* and *b* on line P. The time interval between the two points was deduced from the oscillation (261 Hz) of a tuning-fork (TF) recorded simultaneously. The latency of the response (in milliseconds) was found by counting the number of oscillations recorded between *a* and *b*, irrespective of their length; a constant speed of rotation of the cylinder was not required.

Donders used a large variety of tasks. The stimuli could be lights, colors, written syllables, or spoken syllables, and the responses could be manual key presses or spoken responses. The simplest task was to press a key when a light turned on (or a syllable was spoken). More complex was the go/no-go discrimination task, in which a participant had to press a key (speak a syllable) only when a prespecified target (one of two lights, or a target syllable) was presented. And finally, the most complex, choice response task had lights (syllables) associated with different keys (spoken responses), with the appropriate key to be pressed when the corresponding light went on (or the corresponding syllable was spoken).

Donders' basic, revolutionary observation was that response latencies increased with the complexity of the task: The involvement of more mental stages means more processing time. In passing, he made a number of other seminal observations. "We made the subjects respond with the right hand to the stimulus on the right side, and with the left hand to the stimulus on the left side. When movement of the right hand was required with stimulation on the left side or the other way around, then the time lapse was longer and errors common" (Donders [1868] 1969: 421). This S-R compatibility phenomenon was rediscovered a century later, and came to be called the "Simon effect" (Simon 1967). Furthermore, at the end of his classic article on the measurement of mental processing times, Donders reports that "distraction during the appearance of the stimulus is always punished with prolongation of the process" (Donders [1868] 1969: 428). This observation is interesting in the light of the later research exploiting distraction, such as color-word Stroop (Stroop 1935) and picture-word interference (Lupker 1979).

3. Modeling spoken word recognition

Exactly a century after Donders' seminal article, Morton (1969) published the first modern, discrete two-stage model of word recognition and production, the Logogen model. According to this model, each word is represented by a "logogen", which is a counter collecting

perceptual evidence (during recognition) or conceptual evidence (during production) for the word. When the tally exceeds threshold, the logogen fires, and the syntactic and semantic make-up (during recognition) or articulatory program (during production) of the word is made available. In its original form, assuming a discrete step from word forms to meanings in perceptual processing, the Logogen model no longer gives a correct account of spoken word recognition (although its discreteness assumption may be correct for spoken word production, as I argue later): One of the key observations from modern research of spoken word recognition is that as speech unfolds, multiple word candidates become partially activated and compete for selection (see McQueen, Dahan, and Cutler this volume). The multiple activation concerns not only the forms but also the syntactic properties and meanings of the words. In contrast, the Logogen model holds that only the meaning of the recognized word becomes available.

To account for the activation of multiple lexical candidates, models of spoken word recognition such as the seminal, verbally specified Cohort model of Marslen-Wilson and colleagues (e.g., Marslen-Wilson and Welsh 1978) claim that, on the basis of the first few hundred milliseconds of the speech stream, all words that are compatible with this spoken fragment are activated in parallel in the mental lexicon. For example, when a listener hears the fragment CA, a cohort of words including *cat*, *camel*, *captain* and *captive* becomes activated. Computationally implemented models of spoken word recognition, such as TRACE (McClelland and Elman 1986) and Shortlist (Norris 1994), all instantiate this insight in one form or another.

Evidence for the multiple activation of lexical candidates during word recognition comes from cross-modal semantic priming experiments (e.g., Moss, McCormick, and Tyler 1997; Zwitserlood 1989). For example, Zwitserlood (1989) asked participants to listen to spoken words (e.g., CAPTAIN) or fragments of these words (e.g., CAPT). The participants had to take lexical decisions by means of a key press to written probes that were presented at the offset of the spoken primes. The decision time was measured. The spoken fragments reduced the lexical decision latency for target words that were seman-

tically related to the complete word as well as to cohort competitors. For example, spoken CAPT facilitated the response to the visual probe SHIP (semantically related to *captain*) and also to the probe GUARD (semantically related to *captive*). When the spoken prime was the complete word (CAPTAIN), the lexical decision to SHIP was facilitated but the response to GUARD was not. The activation of multiple meanings was detected as early as 130 milliseconds from the onset of the spoken prime (i.e., during hearing CA), even when the prime was heard in a sentential context that made one of the cohort competitors more plausible than the others.

In activating multiple lexical candidates, the beginning of words plays an important role. Several studies (e.g., Connine, Blasko, and Titone 1993; Marslen-Wilson and Zwitserlood 1989) have shown that when the first phonemes of a spoken non-word prime and the source word from which it is derived differ in more than two phonological features (such as place, voicing, and manner features, e.g., the prime ZANNER derived from MANNER), no cross-modal semantic priming is observed on the lexical decision to a visually presented probe (e.g., STYLE). Marslen-Wilson, Moss, and Van Halen (1996) observed that a difference of one phonological feature between the first phoneme of a *word* prime and its source word leads to no cross-modal semantic priming effect. Using a head-mounted eye-camera to monitor listeners' eye fixations, Allopenna, Magnuson, and Tanenhaus (1998) observed that, for example, hearing the word COLLAR (a rhyme competitor of *dollar*) had less effect than hearing DOLPHIN (a cohort competitor of *dollar*) on the probability of fixating a visually presented target dollar. The reason why Allopenna et al. observed some activation of rhyme competitors while the cross-modal semantic priming studies detected no activation may not only be the use of a different technique (eye tracking vs. cross-modal semantic priming of lexical decision), but also a difference in what was measured. Whereas Allopenna et al. measured the activation of the rhyme competitor directly (i.e., the effect of auditorily presenting COLLAR on the activation of *dollar*), the cross-modal studies measured rhyme activation indirectly (via the semantic relationship of the rhyme competitor to a test probe). Taken together, the evidence suggests that in spoken

word recognition, cohort competitors are more strongly activated than rhyme competitors, even when the rhyme competitors differ only in the initial phoneme from the actually presented spoken word or non-word.

3.1. The TRACE model

How do computational models account for the time course findings, and is there evidence for top-down feedback during spoken word recognition? For many years, the most prominent implemented model of spoken word recognition has been the TRACE model (e.g., McClelland and Elman 1986). TRACE I was built to model findings on phoneme perception and TRACE II (hereafter TRACE) was developed to specifically address issues in spoken word recognition. TRACE falls into the class of interactive-activation models, with activation feedback from later (i.e., lexical) to earlier (i.e., sublexical) levels in spoken word recognition.

Figure 2 illustrates the architecture of TRACE. There are three layers of nodes in TRACE, which represent word forms: a phonological feature level, a phoneme level, and a word level. Syntactic and semantic levels are not included in the model, and, unlike the Logogen model, the relation to speech production is not specified. TRACE represents time by repeating each node at each level for a great number of time slices. As time progresses, feature nodes are activated in successive time slices. Feature nodes activate phoneme nodes, which in turn activate word nodes. Activation flows upwards as well as downwards, so nodes at previous and upcoming time slices can be activated because of the overlap between features and phonemes and between phonemes and words. In TRACE simulations, feature nodes are activated by mimicking acoustic information at 5-msec intervals, with each phoneme node receiving activation from a span of 11 feature nodes. Each phoneme node activates word nodes and features nodes that are consistent with the phoneme, and each phoneme node inhibits all other phoneme nodes at the same temporal position. Finally, each word node inhibits all other word nodes at the

same temporal position and each word nodes activates all of its constituent phonemes.

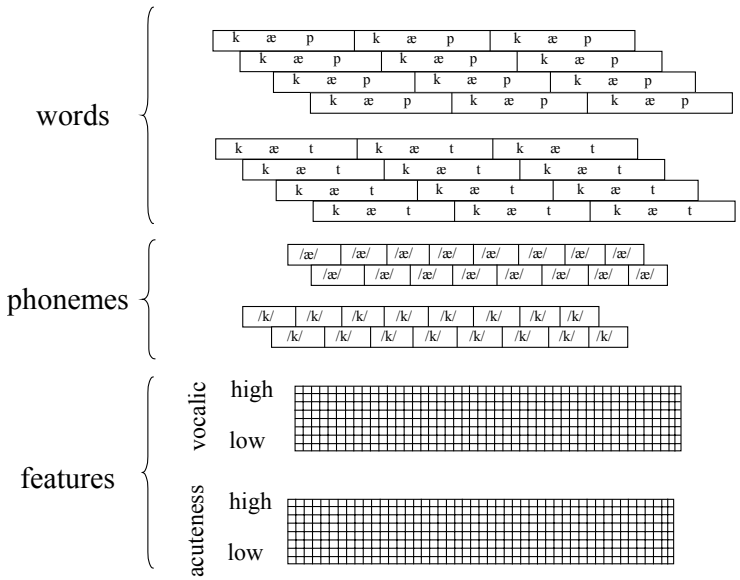

Figure 2. Architecture of the TRACE model of spoken word recognition.

With regard to the time course of word recognition, TRACE largely follows the early Cohort model. Word nodes make up a candidate set (the cohort) as their onset specifications match the acoustic input. When mismatch occurs, word nodes become deactivated, until a single candidate stands out from the rest and is thereby recognized. Words that are activated by their initial phonemes contribute to the activation of their other phonemes by providing top-down feedback, whereas word nodes whose initial phonemes are not matched are suppressed by lateral inhibition and therefore become less activated. Allopenna et al. (1998) showed by TRACE simulations that the model could provide an excellent fit to their findings concerning cohort and rhyme activation using the eye-tracking paradigm.

Logically, spoken word recognition requires bottom-up but not top-down links. Indeed, Frauenfelder and Peeters (1998) ran TRACE

simulations showing that the performance of the model does not worsen when its top-down links are removed. As concerns the empirical side of feedback, TRACE predicts mandatory lexical effects on phoneme processing, which has been challenged by the results of phoneme monitoring experiments. Marslen-Wilson and Warren (1994) examined the effect of lexical status on subphonemic mismatch (i.e., conflicting featural cues as to the identity of a phoneme) by cross-splicing words and non-words. They observed that the lexical status of the source of the cross-spliced material had little effect for words, whereas it had a large effect for non-words. However, in TRACE simulations run by Marslen-Wilson and Warren (1994), the lexical status of the source yielded an effect in the model both for non-words and words, contrary to the real data.

Moreover, in a replication and extension of the study by Marslen-Wilson and Warren (1994), McQueen, Norris, and Cutler (1999) observed that the effect for non-words on phoneme monitoring was dependent on the exact experimental situation. When a wide range of to-be-monitored phonemes was used and the assignment of responses to the left and right hand was varied from trial to trial, as in the study of Marslen-Wilson and Warren (1994), lexical effects in cross-spliced non-words were obtained. However, when the task was made simpler, by using a smaller range of phonemes and keeping the response hand assignment constant, no effect was obtained. Similarly, Cutler et al. (1987) observed that lexical effects in monitoring for word-initial phonemes in monosyllabic targets depended on list composition. In particular, lexical effects were present only when the filler items in the lists varied in the number of syllables. When only monosyllabic fillers were used, no lexical effect was obtained. Moreover, Eimas et al. (1990) observed that lexical effects in phoneme monitoring turned up only when a secondary task directed attention to the lexical level. Lexical effects emerged with noun/verb classification and lexical decision but not with word-length judgment as secondary task, again in contrast to what TRACE predicts.

To conclude, TRACE does a good job in capturing the overall time course of word recognition. However, there is little supporting evidence for the mandatory top-down feedback implemented in

TRACE. Given the success of the model in capturing the time course findings, it is important to know whether similar modeling approaches without feedback can be more successful. Norris (1994) and Norris et al. (2000a) claim that Shortlist presents such an approach.

3.2. The Shortlist model

Shortlist, developed by Norris (1994), combines a recurrent phoneme recognition network and a lexical competition network, in which words detected in the input speech stream are entered as candidates (the shortlist) and compete with each other for recognition (see Figure 3). Syntactic and semantic levels are not included in the model, and the relation to speech production is not specified. The lexical competition network of Shortlist is roughly equivalent to the word level of TRACE but with words included only once and all subthreshold-activated words and their connections removed. Between-word inhibition in Shortlist's competition network is proportional to the phonological overlap between the words. The competition network is wired on the fly on the basis of the phonemes detected by the phoneme recognition network. In the actual simulations, phoneme strings are looked up serially in an electronic dictionary to make simulations with a realistic vocabulary feasible. The words in the competition network inhibit each other for a fixed number of processing cycles, after which their activations are recorded. This whole process of looking up words and dynamically wiring them in the competition network is repeated for each subsequent phoneme in the speech signal. The word that stands out from the competition and that best covers all input phonemes is the recognized word.

Shortlist successfully captures the basic findings about the time course of word recognition. Words join the competition network as their onset phonemes match the acoustic input. When mismatch occurs, word nodes become deactivated, until a single candidate stands out from the rest and is thereby recognized. Words whose initial phonemes are not matched are suppressed by lateral inhibition and there-

therefore become less activated. This explains the finding that cohort competitors are more activated than rhyme competitors. Furthermore, Norris et al. (2000a) showed that a variant of Shortlist that was designed to perform phoneme monitoring, Merge, could handle all extant findings that seemingly suggested top-down feedback, leading them to conclude that "feedback is never necessary" (Norris et al. 2000a: 299). Merge's architecture connects input phoneme nodes to lexical nodes, and both types of nodes are connected to phoneme decision nodes. These connections are feedforward-only. Inhibitory competition operates at the lexical level (corresponding to the competition network of Shortlist) and the phoneme decision level.

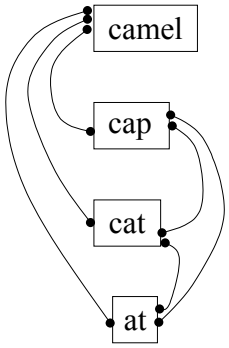

Figure 3. The pattern of inhibitory connections between candidates in a Shortlist competitive lexical network created by presenting as input CA. The figure shows only a subset of candidates matching the input.

Norris et al. (2000a, 2000b) assumed that the connections from the lexical nodes to the phoneme decision nodes in Merge are dynamically built when a listener is asked to perform a phoneme monitoring task (likewise, Shortlist's competition network is built as required). This explains the observations of Cutler et al. (1987), Eimas et al. (1990), and McQueen et al. (1999) that lexical effects in phoneme monitoring are dependent on the exact experimental situation. If the task encourages the use of lexical information, connections are built from both the input phoneme nodes and the lexical nodes to the phoneme decision nodes, which leads to lexical effects on phoneme

monitoring. If the use of lexical knowledge is not encouraged, only the input and phoneme decision nodes are connected, leading to an absence of lexical effects on phoneme monitoring.

In summary, if the speech recognition system also serves speech production, top-down connections should be present and their influence should be detectable. However, to date, there is no convincing positive evidence for top-down feedback in spoken word recognition. Evidence that, on first sight, would seem to suggest top-down influences from the lexical to the sublexical levels, can be explained by a model without feedback (i.e., Shortlist/Merge).

It may be argued that spoken word recognition is a more highly practiced skill than their production: After all, speech recognition precedes production ontogenetically. If the amount of practice is reflected in the strengths of the upward and backward links in a shared recognition/production system, then it may be possible to find evidence for bottom-up, recognition-based feedback (via the stronger recognition links) in production even when there is no evidence for top-down, production-based feedback (via the weaker production links) in recognition. Thus, it is important to see whether there is evidence for feedback in spoken word production, to which I turn now.

4. Modeling spoken word production

We saw that the Logogen model (Morton, 1969) no longer provides a tenable account of spoken word recognition (which involves the activation of multiple meanings rather than a single one, as implied by the Logogen model). However, the model's account of speech production seems to do better: One of the key observations from modern research on speech production is that words are planned in two major steps. In a first, conceptually driven phase, multiple lexical candidates become partially activated and compete for selection. In a second phase, an articulatory program for the highest activated and selected lexical candidate is constructed. There appears to be no form activation for semantic alternatives except for synonyms (Levelt et

al. 1991a, 1999a). The two-step assumption is supported by evidence from both speech errors and chronometric studies (e.g., Levelt et al. 1999a; Roelofs 2003), although it is a hotly debated issue whether the absence of word-form activation for semantic alternatives is due to architectural discreteness (Levelt et al. 1999a 1999b) or to mere functional discreteness (Dell and O'Seaghdha 1991).

Another question that has received much attention is whether there is feedback from phonemes to lexical forms in speech production. One of the classic arguments for feedback is that there are lexical influences on phoneme errors in speech production, the so-called lexical bias. Lexical bias is the finding that form errors create real words rather than non-words with a frequency that is higher than would be expected by chance (e.g., Dell and Reich 1981). Most form errors are non-word errors, but word outcomes tend to be statistically overrepresented. For example, in planning to say "cat", the error "hat" (a word in English) is more likely than the error "zat" (not a word in English). A lexical bias in speech errors is not always observed. While Dell and Reich (1981) found a strong lexical bias in their corpus of errors in spontaneous speech, Garrett (1976) found no such effect and Stemberger (1985) found only a weak effect. In an analysis of the errors in picture naming of fifteen aphasic speakers, Nickels and Howard (1995) found no evidence for lexical bias. A feedback account of the lexical error bias is provided by the DSMSG model.

4.1. The DSMSG model

The DSMSG model (Dell 1986; Dell and O'Seaghdha 1991; Dell et al. 1997) assumes that the mental lexicon is a network that is accessed by spreading activation (the acronym DSMSG was proposed by Dell et al. 1997, and stands for the initials of the authors). Figure 4 illustrates a fragment of the network.

The nodes in the network are linked by equally weighed bidirectional connections. Unlike TRACE and Shortlist, all connections are excitatory; there are no inhibitory links. The network contains nodes

for conceptual features (e.g., ANIMATE, FURRY, etc.), words (e.g., *cat*, *cap*), and phonemes (marked for syllable position, e.g., /onset k/ and /coda t/). The more extensive version of the model proposed by Dell (1986) also contains a level of phonological feature nodes, which are connected to the phoneme nodes. Lexical access starts by supplying a jolt of activation to the set of conceptual features making up the intended thought. Activation then spreads through the network following a linear activation function with a decay factor. Lexical selection is accomplished by selecting the most highly activated word node after a fixed, predetermined number of time steps following the activation of the conceptual feature nodes. Next, the selected word node is given a jolt of activation, and the highest activated onset, nucleus, and coda phonemes are selected after a fixed number of time steps.

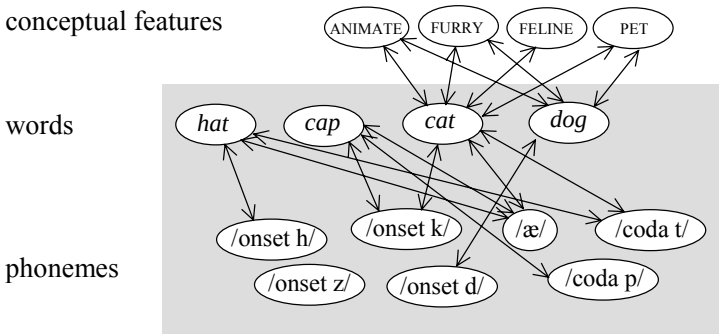

Figure 4. Fragment of the lexical network of the DSMSG model of spoken word production. The shaded area highlights the form level of planning.

The DSMSG model has been specifically designed to account for facts about speech errors of both normal and aphasic speakers: the kind of errors that occur and the constraints on their form and occurrence. According to the model, errors occur when, due to noise in the system, another node than the target one is the most highly activated node and gets erroneously selected. In spite of the presence of backward links in the production network, which might have served speech recognition, Dell et al. (1997) argue for a distinction between

form networks for production and recognition. Under the assumption that word production and recognition are accomplished via one and the same form network, one expects a strong correlation between production and recognition accuracy in aphasia, as verified for the DSMSG model through computer simulations by Dell et al. (1997) and Nickels and Howard (1995). However, such correlations are empirically not observed for form errors by aphasic speakers (e.g., Dell et al. 1997; Nickels and Howard 1995). A distinction between input-form and output-form networks would explain the dissociations between production and recognition capabilities observed in aphasia.

Due to the DSMSG model's interactive nature, semantic and form activation closely follow each other and overlap in time. The model accounts for the finding that semantic activation *precedes* form activation by assuming that the semantic and form effects reflect the timing of the jolts of activation given to the network (with the jolt to the conceptual features preceding the jolt for word-form encoding) rather than activation spreading within the network itself (Dell and O'Seaghdha 1991). The lexical error bias is explained as due to backward spreading of activation from shared phoneme nodes to word nodes (e.g., from the /æ/ node activated by the target *cat* back to *cat* and the competitors *cap* and *hat*) and from these word nodes to other phoneme nodes (i.e., from *hat* to /onset h/). This does not happen for non-words, because there are no word nodes for such items in the network (i.e., there is no node *zat* to activate /onset z/). Thus, it is more likely that in planning to say "cat", /onset h/ is erroneously selected (yielding the error "hat") than that /onset z/ is selected (yielding the error "zat"). In the DSMSG model, activation spreads back automatically from phoneme nodes to word nodes. Thus, as in TRACE, lexical influences on phoneme processing in the DSMSG model are mandatory.

Similar to the lexical effects on phoneme processing in spoken word recognition, and contrary to what the interactive account of the DSMSG model implies, however, the lexical error bias is not a mandatory effect, as already suggested by the seminal study of Baars, Motley, and MacKay (1975). That the lexical error bias is not an inevitable effect is also suggested by the absence of the bias in a num-

ber of error corpora. Baars et al. observed that when all the target and filler items in an error-elicitation experiment are non-words, word slips do not exceed chance. Only when some words are included in the experiment as filler items does the lexical error bias appear. This effect of the filler context should not occur with automatic backward spreading of activation. Therefore, Levelt (1989) and Levelt et al. (1999a), among others, have argued that lexical bias is not due to production-internal activation feedback but that the error bias is at least partly due to self-monitoring of speech planning by speakers. When an experimental task exclusively deals with non-words, speakers do not bother to attend to the lexical status of their speech plan (as they normally often do, apparently), and lexical bias does not arise. Levelt (1989) proposed that self-monitoring of speech planning and production is achieved through the speaker's speech comprehension system, and this assumption has also been adopted for WEAVER++ (Levelt et al. 1999a).

4.2. The WEAVER++ model

WEAVER++ (Levelt et al. 1999a; Roelofs 1992, 1996, 1997a, 1997b, 1998, 1999, 2003; Roelofs and Meyer 1998) assumes that word planning is a staged process, moving from conceptual preparation (including the conceptual identification of a pictured object in picture naming), via lemma retrieval (recovering the word as syntactic entity, including its syntactic properties, crucial for the use of the word in phrases and sentences) to word-form encoding, as illustrated in Figure 5.

Unlike the DSMSG model, WEAVER++ assumes two different lexical levels, namely levels of lemmas and morphemes (the latter representations are involved in word-form encoding), but this is not important for present purposes (see Levelt et al. 1999a, and Roelofs, Meyer, and Levelt 1998, for a theoretical and empirical motivation of the distinction). Comprehending spoken words traverses from word-form perception to lemma retrieval and conceptual identification. In the model, concepts and lemmas are shared between production and

comprehension, whereas there are separate input and output representations of word forms. Consequently, the flow of information between the conceptual and the lemma stratum is bidirectional (Roelofs 1992), whereas it is unidirectional between lemmas and forms as well as within the form strata themselves (top-down for production and bottom-up for comprehension, like in Shortlist). After lemma retrieval in production, spoken word planning is a strictly feedforward process (Roelofs 1997a). Similar to what is assumed in the Logogen model, the transition from lexical selection to word-form encoding in WEAVER++ is a discrete step in that only the form corresponding to a selected lemma becomes activated.

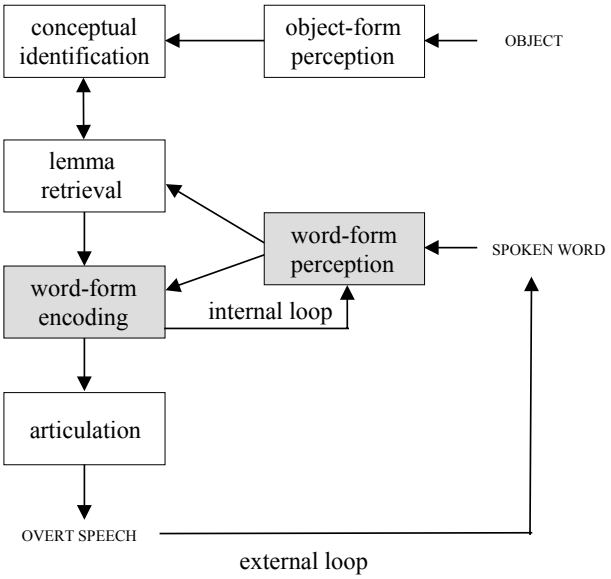

Figure 5. Flow of information in the WEAVER++ model during object naming and spoken word recognition. The shaded boxes indicate the form levels of recognition and production.

Following Levelt (1989), the WEAVER++ model incorporates two self-monitoring loops, an internal and an external one, both operating via the speech comprehension system. Functional brain imag-

ing studies also have suggested that self-monitoring and speech recognition are served by the same neural structures (e.g., McGuire, Silbersweig, and Frith 1996; Paus et al. 1996). The external loop involves listening to self-produced overt speech, whereas the internal loop (which is assumed to be partly responsible for error biases) includes monitoring the speech plan by feeding a rightward incrementally generated phonological word back into the speech comprehension system (Levelt et al. 1999a). A phonological word representation specifies the syllables and, for polysyllabic words, the stress pattern across syllables. Thus, in WEAVER++ there exists feedback of activation from phonemes to lexical forms (see Levelt et al. 1999a, 1999b, for an extensive discussion of this point), except that the feedback engages the speech comprehension system rather than the production system itself. Form production and recognition are achieved by separate but closely linked feedforward systems.

Word planning in WEAVER++ is supported by a lexical network. There are three network strata, shown in Figure 6. A conceptual stratum represents concepts as nodes and links in a semantic network. A syntactic stratum contains lemma nodes, such as *cat*, which are connected to nodes for their syntactic class (e.g., *cat* is a noun, N). And a word-form stratum represents morphemes, phonemes, and syllable programs. The form of monosyllables such as *cat* establishes the simplest case with one morpheme <cat>, phonemes such as /k/, /æ/, and /t/, and one syllable program [kæt], specifying the articulatory gestures. Polysyllabic words such as *tiger* have their phonemes connected to more than one syllable program; for *tiger*, these program nodes are [taɪ] and [gər]. Polymorphemic words such as *catwalk* have one lemma connected to more than one morpheme; for *catwalk* these morphemes are <cat> and <walk>.

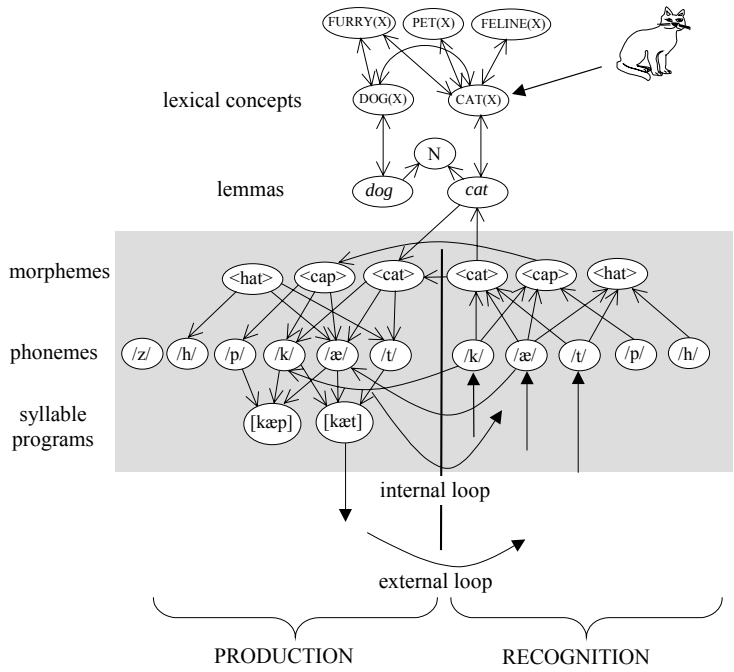

Figure 6. Fragment of the word production and comprehension networks of the WEAVER++ model. The shaded area highlights the form levels of recognition and production.

Information needed for word production planning is retrieved from the network by spreading activation. For example, a perceived object (e.g., cat) activates the corresponding concept node (i.e., CAT(X)). Activation then spreads through the network following a linear activation rule with a decay factor (cf. DSMSG). Each node sends a proportion of its activation to the nodes it is connected to. As in the DSMSG model, there are no inhibitory links. For example, CAT(X) sends activation to other concepts such as DOG(X) and to its lemma node *cat*. Selection of nodes is accomplished by production rules. A production rule specifies a condition to be satisfied and an action to be taken when the condition is met. A lemma retrieval

production rule selects a lemma if the connected concept is flagged as the goal concept. For example, *cat* is selected for CAT(X) if it is the goal concept and *cat* has reached a critical difference in activation compared to other lemmas. The actual moment in time of firing of the production rule is determined by the ratio of activation of the lemma node and the sum of the activations of the other lemma nodes.

A selected lemma is flagged as the goal lemma. A morphological production rule selects the morpheme nodes that are connected to the selected lemma (<cat> is selected for *cat*). Phonological production rules select the phonemes that are connected to the selected morphemes (/k/, /æ/, and /t/ for <cat>) and rightward incrementally syllabify the phonemes (e.g., /k/ is made syllable onset: onset(/k/)) to create a phonological word representation. Finally, phonetic production rules select syllable-based motor programs that are appropriately connected to the syllabified phonemes (i.e., [kæt] is selected for onset(/k/), nucleus(/æ/), and coda(/t/)). The moment of selection of a program node is given by the ratio of activation of the target node and the sum of the activations of all other program nodes.

WEAVER++ implements a number of specific claims about how the spoken word production and recognition networks are related, as shown in Figure 6. To account for interference and facilitation effects from auditorily presented distractor words on picture naming latencies, Roelofs (1992, 1997a; Roelofs, Meyer, and Levelt 1996) assumed that form information activated in a speech recognition network activates compatible phoneme, morpheme, and lemma representations in the production network (see also Levelt et al. 1999a). For convenience, Figure 6 shows phoneme and lexical form nodes in the comprehension network (following McClelland and Elman 1986; Norris 1994), but this is not critical for present purposes (see Lahiri and Marslen-Wilson 1991, for a model of speech recognition that has no such phonemes). Covert self-monitoring involves feeding the rightward incrementally constructed phonological word representation from speech production into the speech comprehension system (Levelt et al. 1999a).

A lexical error bias arises within WEAVER++ in at least three ways (for discussion, see Roelofs in press). First, the bias occurs

when speakers employ lexicality as an explicit monitoring criterion, as suggested by Levelt and colleagues (1989; Levelt et al. 1999a). Second, lexical bias arises as some form-related errors are due to lemma or morpheme selection failures (i.e., if they are "malapropisms") rather than phoneme selection failures. On the classic feedback account (e.g., Dell and Reich 1981), lexical bias arises in phoneme selection, but this does not need to be the case. Some errors may be lexical errors, perfectly in line with a theory that assumes a feedforward-only relation between lexical forms and their phonemes. A malapropism may occur when a speaker can generate only an incomplete form representation of the intended word, as in a tip-of-the-tongue state. This incomplete form is fed back to the conceptual system via the comprehension system, which leads to the activation of the lemmas of words that are phonologically related to the target. These lemmas typically will be semantically unrelated to the target. If one of these lemmas of the appropriate grammatical category is selected, a malapropism will occur. Third, lexical bias occurs in accessing motor programs for syllables (i.e., syllable program nodes in the network). Because the feedback loop through the speech comprehension system activates compatible morpheme representations in the production network, which activate corresponding syllable program nodes, the loop favors the selection of syllable programs that correspond to words. Note that in a context that de-emphasizes self-monitoring and a lexical involvement, such as the all-nonwords condition of Baars et al. (1975), a lexical-error bias should not occur, in agreement with the empirical findings. To conclude, in a model without production-internal backward links from phonemes to lexical forms such as WEAVER++, there are several factors that give rise to a tendency to produce word over non-word errors at a higher rate than chance.

The WEAVER++ model accounts for the finding that semantic effects precede form effects in time in terms of network activation patterns during the successive planning stages of lemma retrieval and word-form encoding. The assignment of the semantic and form effects of spoken distractors in object naming to different planning levels is independently supported by the finding that spoken cohort and

rhyme distractors yield facilitation effects of similar size in picture naming (Collins and Ellis 1992; Meyer and Schriefers 1991; Meyer and Van der Meulen 2000), whereas they yield differential effects in spoken word recognition tasks. Cohort competitors are more strongly activated in spoken word comprehension than rhyme competitors, even when the first phoneme of the rhyme competitor deviates by only two phonological features from the actually presented spoken word (e.g., Allopenna et al. 1998; Connine et al. 1993; Marslen-Wilson and Zwitserlood 1989). In the next section, I argue that the dissociation of cohort and rhyme effects between production and recognition supports the assignment of semantic and form effects of spoken distractors to different planning levels in production (contrary to Starreveld and La Heij 1996) and that the dissociation challenges a shared production/recognition system.

5. Cohort versus rhyme effects

Meyer and Schriefers (1991) observed that when spoken cohort or rhyme distractors are presented over headphones during the planning of monosyllabic picture names (e.g., the spoken distractors CAP or HAT during planning to say the target word "cat"), both distractors yield faster latencies compared to unrelated distractors. When cohort or rhyme distractors (e.g., METAL or VILLAIN) are auditorily presented during the planning of disyllabic picture names (e.g., *melon*), both distractors yield faster latencies too. When the difference in time between distractor and target presentation is manipulated (e.g., SOA = -300, -150, 0, 150 ms), the SOA at which the faster latencies are first detected differs between cohort and rhyme distractors. In particular, faster latencies occur at an earlier SOA for cohort than for rhyme distractors (i.e., respectively, SOA = -150 ms and SOA = 0 ms). At SOAs where both effects are present (i.e., 0 and 150 ms), the magnitude of the facilitation effect from cohort and rhyme distractors was the same in the study of Meyer and Schriefers (1991). Collins and Ellis (1992) and Meyer and Van der Meulen (2000) made similar observations. Moreover, Meyer and Van der Meulen (2000) observed

analogous effects of cohort and rhyme distractors on speakers' eye-gaze durations. Earlier studies by Meyer and colleagues using an eye-tracker to measure gaze durations during the naming of objects (e.g., Meyer et al. 1998) showed that a speaker keeps fixating a to-be-named perceived object until the phonological form of the object name has been prepared. Meyer and Van der Meulen (2000) observed that the eye-gaze durations were shortened to an equal extent with cohort and rhyme distractors as compared to unrelated distractor words. In contrast, in spoken word recognition, cohort competitors are more strongly activated than rhyme competitors.

The difference between the findings from cross-modal studies in the spoken word recognition literature and the findings from spoken distractors in picture naming is readily explained if one assumes that spoken distractor words do not activate rhyme competitors at the lemma level, but that rhyme relatedness effects result from activation of the corresponding phonemes in the production lexicon. Roelofs (1997a) provides such an account, implemented in WEAVER++, and reports computer simulations of the effects. On this account, METAL and VILLAIN activate the production phonemes that are shared with *melon* to the same extent (respectively, the phonemes of the first and second syllable), which explains the findings on picture naming of Meyer and Schriefers (1991). At the same time, METAL activates the lemma of *melon* whereas VILLAIN does not, which explains the findings on spoken word recognition. WEAVER++ simulations have shown that cohort activation does not result in facilitation of lemma retrieval in the model, unless there is also a semantic relationship involved (cf. Levelt et al. 1999b), as with *cat* and *camel*.

The finding that spoken cohort and rhyme distractors yield facilitation effects of similar size in picture naming, whereas they yield differential effects in spoken word recognition tasks challenges a shared production/recognition system. If the system is shared, it seems difficult to explain why priming the second syllable of a picture name by a spoken rhyme distractor leads to the same amount of facilitation as priming the name's first syllable by a spoken cohort distractor, while cohort competitors are more strongly activated than rhyme competitors in spoken word recognition. In contrast, if form

recognition and production are achieved by separate feedforward systems (as in WEAVER++), then activation of forms in the recognition system may yield differential activation of cohort and rhyme competitors at the lexical level, while the corresponding cohort and rhyme phonemes may be equally activated in the production lexicon. If there is no feedback in the production lexicon, equal activation of cohort and rhyme phonemes does not lead to differential activation of cohort and rhyme competitors at the lemma level in production. This explains the differential influence of serial order on lexical and sublexical levels in production and recognition. The account requires separate form production and recognition networks without feedback.

6. Speaking while hearing words

Although in interactive models like TRACE and DSMSG feedback occurs automatically, there is the logical possibility that in a single recognition/production system, the bottom-up links may be operative only during actual word recognition and the top-down links only during actual word production. To evaluate this possibility, evidence from combined recognition/production tasks rather than from pure recognition or pure production tasks is critical. In the previous section, I discussed evidence from one task that meets the production/recognition simultaneity condition, namely auditory picture-word interference, which did not support feedback. In this section, evidence from two other tasks is discussed, namely auditory lexical decision during object naming and combined auditory priming/speech preparation.

6.1. Auditory lexical decision during object naming

Levelt et al. (1991a) combined picture naming with auditory lexical decision. Participants were asked to name pictured objects and, on some critical trials, they had to interrupt the preparation of the pic-

ture name and to make a lexical decision by means of a key press to an auditory probe presented after picture onset (i.e., with SOAs of 73, 373, or 673 ms). Thus, the speakers had to monitor for the lexical status of spoken probes while preparing to say the name of the object. The auditory lexical decision latency was the dependent variable. In this double-task situation, both forward and backward links should be operative, to meet the recognition (auditory lexical decision) and production (object naming) requirements of the double task. Thus, the double task meets the simultaneity condition for obtaining evidence for feedback, if it exists.

In one experiment, Levelt et al. looked at the time course of semantic and phonological effects. At the early SOA (73 ms), the lexical decision latencies were slowed down for spoken probes semantically related to the picture name as compared with unrelated probes. For example, in planning the production of "cat", lexical decisions were slower for the spoken probe DOG than for the probe CHAIR. In contrast, at all SOAs interference was obtained for phonologically related probes compared with unrelated ones. For example, decision latencies were longer for CAP than for CHAIR in planning the production of "cat" at all SOAs. The finding that the semantic effect was confined to the early SOA suggests that there is no feedback from phonemes to words in the speech production system, contrary to what is held by the DSMSG model. As we saw, the DSMSG model accounts for the timing of latency effects by assuming that the effects reflect the timing of jolts of activation to the conceptual and lexical representations rather than the activation within the network itself. However, although this may explain the early semantic effect, it fails to explain why phonological effects occur both early and late in time (whereas the jolt for word-form encoding is given only once, after lexical selection).

Another experiment conducted by Levelt et al. tested for phonological activation of semantic alternatives to the target. According to the DSMSG model, the phonemes of semantic competitors of the target (e.g., *dog* as a competitor of *cat*) should become active, whereas according to WEAVER++, they should not. Levelt et al. obtained no effect on the lexical decision latencies for spoken probes that were

that were phonologically related to semantic competitors (LOG), whereas they did obtain semantic interference for such semantic competitors (DOG) themselves. This result supports the discrete view.

However, in a reply to Levelt et al., Dell and O'Seaghdha (1991) presented the results of computer simulations with the DSMSG model that suggested that phoneme activation does not necessarily happen for words that are phonologically related to semantic competitors (LOG). Because the phonemes are only indirectly activated (through the conceptual features shared between *cat* and *dog* and the phonemes shared between *dog* and *log*), they are not much activated at all in the DSMSG model, in agreement with the empirical findings. Thus, even though the DSMSG model is not architecturally discrete (as WEAVER++), it behaves in a functionally discrete manner. A problem with this counter-argument by Dell and O'Seaghdha is that it is based on activation patterns in the DSMSG network occurring *without* auditorily presented probes (see Levelt et al. 1991b, for discussion), which is not the situation tested by Levelt et al. As mentioned earlier, the influence of feedback is presumably best felt in combined production/recognition tasks. However, Dell and O'Seaghdha did not put the DSMSG model to such a test, thereby reducing the effect of the feedback links present in the model.

Another argument against the conclusions of Levelt et al. (1991a) came from Harley (1993), who presented the results from simulations using a very different network model, which, similar to TRACE, contained inhibitory links between nodes representing incompatible information (e.g., word nodes inhibited each other). In this model, the phonemes of phonological relatives of semantic alternatives did not become much activated, again, in simulations without auditory distractors. Furthermore, despite the backward spreading of activation in the network, there was no late semantic rebound (i.e., the model exhibited only an early semantic effect). Relevant for the issue of shared versus separate recognition/production systems, however, the semantic effect occurred early in Harley's model because there are no backward links from words to their meanings. Thus, the simulations cannot be taken as evidence for a single system achieving both production and recognition (word comprehension requires

links from word forms to meanings).

To conclude, there is no positive evidence for feedback from a double auditory lexical decision/picture naming task. Of course, one may argue that the critical task was a perceptual one (i.e., auditory lexical decision), so the feedback effect should have come from the supposedly weaker speech production links. An experiment that engages the supposedly stronger recognition-based feedback links in a production task may provide a stronger test, which I discuss next.

6.2. Combined auditory priming/speech preparation

In Roelofs (2002), I report a study that tested for the combined effect of preparing the early parts of a to-be-produced word and auditory priming of later parts of that word. Participants produced disyllabic words out of small response sets in reaction to prompts. The words were unrelated or shared the first syllable (e.g., the syllable *me* in the responses *melon*, *metal*, and *merit*), which allowed for preparation of that syllable. At prompt onset, auditory syllable primes were presented that matched the second syllable of the response or not (e.g., LON or TAL for *melon*). Note that in this task situation, again, both forward and backward links should be operative, because production and recognition are involved. Thus, the combined auditory priming/speech preparation task meets the production/recognition simultaneity condition for obtaining evidence for feedback, if it exists.

Because preparation and priming aimed at different serial loci (i.e., in the example, the first and second syllable of the target), their combined effect should be additive or interactive depending on the theoretical position. Under the assumption of feedback from phonemes to lexical forms in production, the auditory second-syllable prime LON should facilitate the production of "melon" both directly and indirectly. The auditory prime LON activates the phonemes /l/, /ə/, and /n/ in the network (direct priming of the second syllable), which may spread activation back to the word node *melon*, which in turn may forwardly activate /m/ and /e/ (indirect priming of the first syllable). Such indirect priming is not possible when the first syllable

is already prepared. Thus, the feedback view predicts an interaction between priming and preparation: The size of the effect of second syllable priming should depend on whether or not the first syllable is already prepared. However, the experiment yielded effects of both priming and preparation, but there was not even a hint of an interaction, challenging the assumption of feedback links in the production form network. In contrast, WEAVER++ simulations showed that this model could account for the observed, perfectly additive effects.

To conclude, there is no positive evidence for feedback from chronometric tasks that involve both spoken word production and recognition. Of course, although most computationally implemented models of production and recognition have addressed latency findings, other evidence cannot be ignored. In the introduction section, I mentioned that at the time Donders conducted his chronometric studies, Wernicke (1874) proposed that the left posterior superior temporal lobe is involved in both speech recognition and production. However, this brain area is broad and it is typically not uniformly affected by damage. Therefore, an important issue is whether exactly the same area is involved in recognition and production, as Wernicke claimed, or whether recognition and production are supported by distinct subregions of the left posterior superior temporal area.

7. Co-activation in functional neuroimaging studies

Neuroimaging studies have confirmed, independently for production and recognition, Wernicke's observation that the posterior superior temporal lobe is involved in production and recognition. In particular, studies suggest that this area in both hemispheres is involved in speech recognition and that the area in the left hemisphere is involved in speech production (see Buchsbaum et al. 2001, and Hickok and Poeppel 2000, for reviews).

It has been observed that transcranial magnetic stimulation of the posterior superior temporal lobe in either hemisphere disrupts speech perception. Furthermore, single unit recordings during brain surgery have revealed cells in the posterior superior temporal area in both

hemispheres that respond selectively to speech input. Moreover, pure word deafness is commonly associated with bilateral lesions involving the area. Although there is bilateral involvement of the posterior superior temporal lobe in speech recognition, the speech stream seems to be asymmetrically analyzed in the time domain, with the right hemisphere analyzing phonetic information over a longer time window (i.e., 150-250 ms) than the left hemisphere (25-50 ms). Furthermore, several studies (see Buchsbaum et al. 2001, for a review) have suggested an involvement of the *left* posterior superior temporal lobe in speech production. For example, picture naming is facilitated by transcranial magnetic stimulation of the area in the left but not in the right hemisphere. Furthermore, a meta-analysis of 58 functional imaging studies by Indefrey and Levelt (2000) revealed activation of the left posterior superior temporal area in object naming, word generation, and syllable rehearsal.

Recently, Buchsbaum et al. (2001) conducted an event-related fMRI study to determine to what extent the posterior superior temporal lobe is involved in both speech recognition and production using a task that had both production and recognition components. While undergoing fMRI, participants listened to three non-words presented at a rate of one per second, which then had to be silently rehearsed for 27 seconds.

In relation to the perceptual phase of a trial, Buchsbaum et al. observed bilateral activation of the primary auditory cortex (i.e., Heschl's gyrus) and adjacent areas, and also activation of some frontal and parietal areas. Related to the motor phase of a trial, they observed, predominantly for the left hemisphere, activation of the lateral premotor and inferior frontal cortex, and also activation of some temporal areas. Most importantly, activation related to both the perceptual and motor phases of a trial was observed for two posterior superior temporal regions. One region concerned the superior temporal sulcus and lateral posterior superior temporal gyrus (pSTG), henceforth the ventral site. The other region concerned the posterior superior temporal planum (pSTP) and parietal operculum (PO), henceforth the dorsal site. Some participants showed activation of the ventral site in the right hemisphere, but no participant showed activa-

tion of the dorsal site in the right hemisphere. Less relevant for now, activation related to both the perceptual and motor phases of a trial was also observed for lateral premotor and inferior frontal regions, roughly corresponding to Broca's area (Brodmann's areas 44 and 45). Figure 7 illustrates the relevant perisylvian areas by means of a lateral view of the left hemisphere with the areas inside the Sylvian fissure exposed.

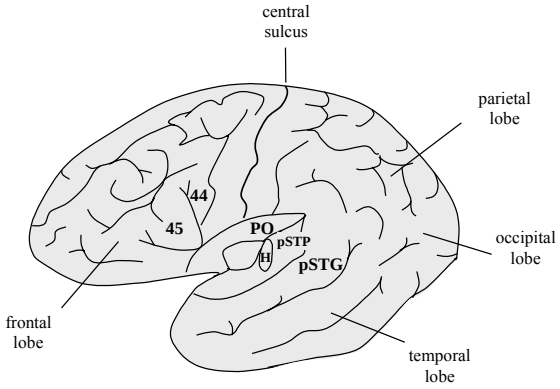

Figure 7. Lateral view of the left hemisphere of the human brain with the areas inside the Sylvian fissure exposed: pSTG = posterior superior temporal gyrus; pSTP = posterior superior temporal planum; PO = parietal operculum; H = Heschl's gyrus, which contains the primary auditory cortex; 44 and 45 refer to Brodmann's designations and make up Broca's area.

There were, however, differences in the time course of activation of the two posterior superior temporal sites. The ventral site showed more robust perception-related activation than did the dorsal site. Conversely, the dorsal site showed more robust motor-related activation than did the ventral site. The top panel of Figure 8 shows the observed blood flow responses in the ventral and dorsal subregions of Wernicke's area.

On the basis of these results, Buchsbaum et al. (2001) argued that there is overlap in the neural systems that participate in phonological aspects of speech recognition and production, supporting models (like Wernicke's) that posit overlap in the phonological input and output systems. However, on the basis of the results, it seems diffi-

cult to distinguish between overlapping systems and closely linked ones. Because the form input and output networks in WEAVER++ are tightly connected, and activation of one form network automatically leads to the activation of the other, the type of co-activation observed by Buchsbaum et al. (2001) is entailed. This claim was supported by WEAVER++ simulations, assuming that cerebral blood flow is a gamma function of network activation (cf. Roelofs and Hagoort 2002) and that the form-perception network is associated with ventral Wernicke and the form-production network with dorsal Wernicke. The latter is supported by anatomical evidence showing that Heschl's gyrus (primary auditory cortex) is connected via monosynaptic pathways with ventral Wernicke, but there are no direct connections with dorsal Wernicke (Wise et al. 2001).

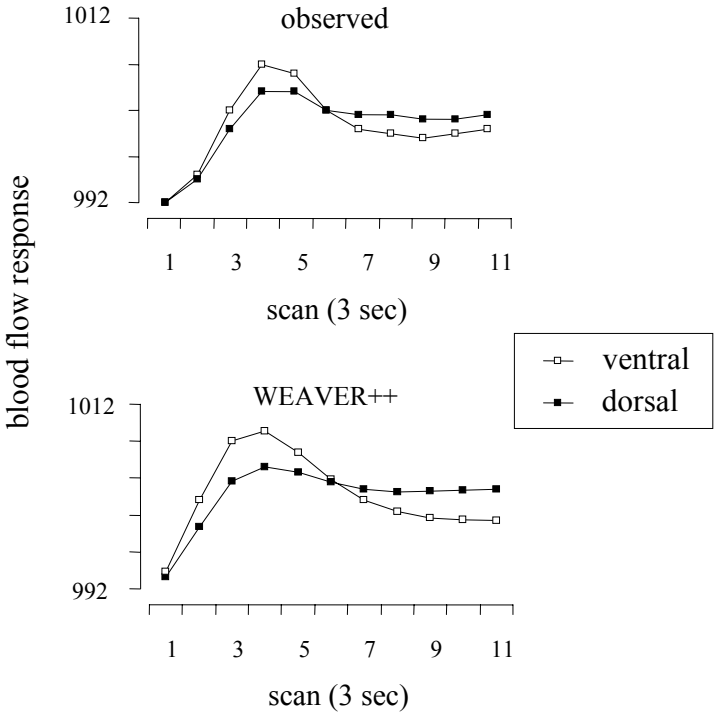

Figure 8. Blood flow responses in the ventral and dorsal subregions of Wernicke's area observed in the fMRI study by Buchsbaum et al. (2001) and in WEAVER++ simulations.

The bottom panel of Figure 8 shows the simulated blood flow responses in WEAVER++. During speech recognition, form activation in the recognition network automatically activates the corresponding forms in the production network. Because the activation of the production network is indirect, it will be less than during actual speech production. Similarly, during production, form activation in the production network automatically activates the corresponding forms in the recognition network. Again, because the activation of the recognition network is indirect, it will be less than during actual speech recognition. Figure 8 shows that the simulated blood flow responses and their dependence on the task (production versus perception) are in agreement with the brain's blood flow responses observed by Buchsbaum et al. (2001).

To conclude, Wernicke may have been right in assuming that the left posterior superior temporal lobe is involved in both speech recognition and production. However, there is no conclusive evidence that exactly the same regions of the left superior temporal area are activated to the same degree in both recognition and production rather than different subregions to different degrees. On the contrary, the latter has received support from functional neuroimaging.

8. Summary and conclusions

Donders' (1868) contributions have now turned 65 twice. During the past century, the study of speech production and recognition has used techniques and has yielded results that would have far exceeded his imagination. However, one of the questions that interested Donders, the relation between speech production and recognition, has received surprisingly little attention. In this chapter, I have made a case for a modern version of Donders' claim that mentally progressing from speech input to output involves a translation process. The case was made primarily on the basis of chronometric findings and their modeling within the recognition and production research traditions.

First, I have argued that there is no conclusive evidence in favor of production-internal or recognition-internal feedback, neither from

recognition tasks nor from production tasks, and not even from tasks that combine both recognition and production. Furthermore, cohorts and rhymes play a different role in production and recognition, which challenges the view of a shared system. Second, although recent functional imaging studies have confirmed the observation of Donders' contemporary Wernicke that there exists a brain area, the posterior superior temporal lobe, which critically participates in both speech recognition and production, there is also evidence that suggests that different subregions of this broad area are differently involved in the two tasks. To conclude, the available evidence supports the idea of separate but closely linked feedforward systems for word-form recognition and production.

References

Allopenna, Paul D., James S. Magnuson and Michael K. Tanenhaus
1998 Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38: 419-439.

Allport, D. Alan
1984 Speech production and comprehension: One lexicon or two? In: Wolfgang Prinz and Andries F. Sanders (eds.), *Cognition and motor processes*, 209-228. Berlin: Springer-Verlag.

Baars, Bernard J., Michael T. Motley and Donald G. MacKay
1975 Output editing for lexical status from artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* 14: 382-39.

Buchsbaum, Bradley R., Gregory Hickok and Colin Humphries
2001 Role of the left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science* 25: 663-678.

Cherry, Colin
1957 *On human communication.* New York: John Wiley.

Caramazza, Alfonso
1997 How many levels of processing are there in lexical access? *Cognitive Neuropsychology* 14: 177-208.

Collins, Alan F. and Andrew Ellis
1992 Phonological priming of lexical retrieval in speech production. *British Journal of Psychology* 83: 375-388.

Connine, Cynthia M., Dawn G. Blasko and Debra Titone
1993 Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language* 32: 193-210.

Cutler, Anne, Jacques Mehler, Dennis Norris and Juan Segui
1987 Phoneme identification and the lexicon. *Cognitive Psychology* 19: 141-177.

Cutler, Anne and Dennis Norris
1988 The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14: 113-121.

Dell, Gary S.
1986 A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93: 283-321.

Dell, Gary S. and Padraig O'Seaghdha
1991 Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). *Psychological Review* 98: 604-614.

Dell, Gary S. and Peter A. Reich
1981 Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior* 20: 611-629.

Dell, Gary S., Myrna F. Schwartz, Nadine Martin, Eleanor M. Saffran and Deborah A. Gagnon
1997 Lexical access in aphasic and nonaphasic speakers. *Psychological Review* 104: 801-838.

Donders, Franciscus C.
1849 De bewegingen der hersenen en de veranderingen der vaatvulling van de Pia Mater, ook bij gesloten onuitzetbaren schedel regtstreeks onderzocht [The movements of the brain and the changes of the content of the Pia Mater, also directly investigated with a closed skull]. *Nederlandsch Lancet. Tijdschrift voor de Geneeskundigen Wetenschappen in Haren Geheelen Omvang* 5: 521-553.

Donders, Franciscus C.
1868 Over de snelheid van psychische processen. *Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868-1869, Tweede reeks* II: 92-120. Reprinted as Donders, Franciscus C. (1969). On the speed of mental processes. *Acta Psychologica* 30: 412-431.

Donders, Franciscus C.
1870 *De physiologie der spraakklanken, in het bijzonder van die der Nederlandsche taal* [The physiology of speech sounds, in particular those of the Dutch language]. Utrecht: Van der Post.

Eimas, Peter D., Suzan B. Marcovitz Hornstein and Paula Patton
1990 Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language* 29: 160-180.

Foss, Donald J.
1969 Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior* 8: 457-462.

Frauenfelder, Uli H. and Guus Peeters
1998 Simulating the time-course of spoken word recognition: An analysis of lexical competition in TRACE. In: Jonathan Grainger and Arthur M. Jacobs (eds.), *Localist connectionist approaches to human cognition*. Hillsdale, NJ: Erlbaum.

Frauenfelder, Uli H. and Juan Segui
1989 Phoneme monitoring and lexical processing: Evidence for associative context effects. *Memory & Cognition* 17: 134-140.

Fromkin, Victoria
1971 The non-anomalous nature of anomalous utterances. *Language* 47: 27-52.

Garrett, Merrill F.
1975 The analysis of sentence production. In: Gordon H. Bower (ed.), *The psychology of learning and motivation*, 133-177. New York: Academic Press.

Glaser, Wilhelm R. and Franz-Josef Düngelhoff
1984 The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance* 10: 640-654.

Griffin, Zenzi M. and Kathryn Bock
2000 What the eyes say about speaking. *Psychological Science* 11: 274-279.

Grosjean, François
1980 Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics* 28: 267-283.

Harley, Trevor A.
1993 Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes* 8: 291-309.

Hickok, Gregory and David Poeppel
2000 Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* 4: 131-138.

Indefrey, Peter and Willem J. M. Levelt
2000 The neural correlates of language production. In: Michael Gazzaniga (ed.), *The new cognitive neurosciences*, 845-865. Cambridge, MA: MIT Press.

Jescheniak, Jörg and Willem J. M. Levelt
1994 Word frequency effects in speech production: Retrieval of syntactic information and phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 824-843.
Lahiri, Aditi and William D. Marslen-Wilson
1991 The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition* 38: 243-294.
Levelt, Willem J. M.
1989 *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
Levelt, Willem J. M.
1999 Models of word production. *Trends in Cognitive Sciences* 3: 223-232.
Levelt, Willem J. M., Ardi Roelofs and Antje S. Meyer
1999a A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22: 1-38.
Levelt, Willem J. M., Ardi Roelofs and Antje S. Meyer
1999b Multiple perspectives on word production. *Behavioral and Brain Sciences* 22: 61-75.
Levelt, Willem J. M., Herbert Schriefers, Dirk Vorberg, Antje S. Meyer, Thomas Pechmann and Jaap Havinga
1991a The time course of lexical access in speech production: A study of picture naming. *Psychological Review* 98: 122-142.
Levelt, Willem J. M., Herbert Schriefers, Dirk Vorberg, Antje S. Meyer, Thomas Pechmann and Jaap Havinga
1991b Normal and deviant lexical processing: A reply to Dell and O'Seaghdha. *Psychological Review* 98: 615-618.
Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler and Michael Studdert-Kennedy
1967 Perception of the speech code. *Psychological Review* 74: 431-461.
Lupker, Stephen J.
1979 The semantic nature of response competition in the picture-word interference task. *Memory & Cognition* 7: 485-495.
MacKay, Donald G.
1987 *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer-Verlag.
MacLeod, Colin M.
1991 Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin* 109: 163-203.
Marslen-Wilson, William
1973 Linguistic structure and speech shadowing at very short latencies. *Nature* 244: 522-523.

Marslen-Wilson, William D., Helen E. Moss and Stef van Halen
1996 Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 22: 1376-1392.
Marslen-Wilson, William D. and Paul Warren
1994 Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review* 101: 653-675.
Marslen-Wilson, William D. and Alan Welsh
1978 Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10: 29-63.
Marslen-Wilson, William D. and Pienie Zwitserlood
1989 Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance* 15: 576-585.
Martin, Nadine, Deborah A. Gagnon, Myrna F. Schwartz, Gary S. Dell and Eleanor M. Saffran
1996 Phonological facilitation of semantic errors in normal and aphasic speakers. *Language and Cognitive Processes* 11: 257-282.
McClelland, James L. and Jeffrey L. Elman
1986 The TRACE model of speech perception. *Cognitive Psychology* 18: 1-86.
McGuire, Philip K., David A. Silbersweig and Chris D. Frith
1996 Functional neuroanatomy of verbal self-monitoring. *Brain* 119: 907-917.
McCusker, Leo X., Michael L. Hillinger and Randolph G. Bias
1981 Phonological recoding and reading. *Psychological Bulletin* 89: 217-245.
McQueen, James M., Dennis Norris and Anne Cutler
1999 Lexical influences in phonetic decision-making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance* 25: 1363-1389.
Meyer, Antje S. and Herbert Schriefers
1991 Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17: 1146-1160.
Meyer, Antje S., Astrid M. Sleiderink and Willem J. M. Levelt
1998 Viewing and naming objects: Eye movements during noun phrase production. *Cognition* 66: B25-B33.

Meyer, Antje S. and Femke F. van der Meulen
2000 Phonological priming effects on speech onset latencies and viewing times in object naming. *Psychonomic Bulletin & Review* 7: 314-319.
Monsell, Stephen
1987 On the relation between lexical input and output pathways for speech. In: Alan Allport, Donald G. MacKay, Wolfgang Prinz and Eckart Scheerer (eds.), *Language perception and production: Relationships between listening, speaking, reading, and writing*, 273-311. London: Academic Press.
Morton, John
1969 Interaction of information in word recognition. *Psychological Review* 76: 165-178.
Moss, Helen E., Samantha F. McCormick and Lorraine K. Tyler
1997 The time course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes* 12: 695-731.
Nickels, Lyndsey and David Howard
1995 Phonological errors in aphasic naming: Comprehension, monitoring, and lexicality. *Cortex* 31: 209-237.
Norris, Dennis
1994 Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52: 189-234.
Norris, Dennis, James M. McQueen and Anne Cutler
2000a Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* 23: 299-325.
Norris, Dennis, James M. McQueen and Anne Cutler
2000b Feedback on feedback on feedback: It's feedforward. *Behavioral and Brain Sciences* 23: 352-370.
Paus, Tomas, David W. Perry, Robert J. Zatorre, Keith J. Worsley and Alan C. Evans
1996 Modulation of cerebral blood flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. *European Journal of Neuroscience* 8: 2236-2246.
Price, Cathy, Peter Indefrey and Miranda van Turennout
1999 The neural architecture underlying the processing of written and spoken word forms. In: Colin M. Brown and Peter Hagoort (eds.), *The neurocognition of language*, 212-240. Oxford: Oxford University Press.
Roelofs, Ardi
1992 A spreading-activation theory of lemma retrieval in speaking. *Cognition* 42: 107-142.

Roelofs, Ardi
1996 Serial order in planning the production of successive morphemes of a word. *Journal of Memory and Language* 35: 854-876.
Roelofs, Ardi
1997a The WEAVER model of word-form encoding in speech production. *Cognition* 64: 249-284.
Roelofs, Ardi
1997b Syllabification in speech production: Evaluation of WEAVER. *Language and Cognitive Processes* 12: 657-693.
Roelofs, Ardi
1998 Rightward incrementality in encoding simple phrasal forms in speech production: Verb-particle combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24: 904-921.
Roelofs, Ardi
1999 Phonological segments and features as planning units in speech production. *Language and Cognitive Processes* 14: 173-200.
Roelofs, Ardi
2002 Spoken language planning and the initiation of articulation. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology* 55: 465-483.
Roelofs, Ardi
2003 Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review* 110: 88-125.
Roelofs, Ardi
in press Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). *Psychological Review*.
Roelofs, Ardi and Peter Hagoort
2002 Control of language use: Cognitive modeling of the hemodynamics of Stroop task performance. *Cognitive Brain Research* 15: 85-97.
Roelofs, Ardi and Antje S. Meyer
1998 Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24: 922-939.
Roelofs, Ardi, Antje S. Meyer and Willem J. M. Levelt
1996 Interaction between semantic and orthographic factors in conceptually driven naming: Comment on Starreveld and La Heij (1995). *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 246-251.
Roelofs, Ardi, Antje S. Meyer and Willem J. M. Levelt
1998 A case for the lemma-lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition* 69: 219-230.

Schriefers, Herbert, Antje S. Meyer and Willem J. M. Levelt
1990 Exploring the time-course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language,* 29: 86-102.
Shattuck-Hufnagel, Stefanie
1979 Speech errors as evidence for a serial-order mechanism in sentence production. In: William E. Cooper and Edward C. T. Walker (eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*, 295-342. Hillsdale, NJ: Erlbaum.
Simon, J. Richard
1967 Choice reaction time as a function of auditory S-R correspondence, age and sex. *Ergonomics* 10: 659-664.
Starreveld, Peter A. and Wido La Heij
1996 Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 896-918.
Stemberger, Joseph P.
1985 An interactive activation model of language production. In: Andrew W. Ellis (ed.), *Progress in the psychology of language*, 143-186. London: LEA.
Stroop, J. Ridley
1935 Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18: 643-662.
Swinney, David, William Onifer, Penny Prather and Max Hirshkowitz
1979 Semantic facilitation across modalities in the processing of individual words and sentences. *Memory & Cognition* 7: 159-165.
Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard and Julie C. Sedivy
1995 Integration of visual and linguistic information during spoken language comprehension. *Science* 268: 1632-1634.
Warren, Robert E.
1972 Stimulus encoding and memory. *Journal of Experimental Psychology* 94: 90-100.
Wernicke, Carl
1874 *Der aphasische Symptomenkomplex*. Breslau: Cohn & Weigert.
Wise, Richard J.S., Sophie K. Scott, S. Catrin Blank, Cath J. Mummery, Kevin Murphy and Elizabeth A. Warburton
2001 Separate neural subsystems within 'Wernicke's area'. *Brain* 124: 83-95.
Zwitserlood, Pienie
1989 The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition* 32: 25-64.