Spoken word production: A theory of lexical access

Willem J. M. Levelt*

Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH, Nijmegen, The Netherlands

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on May 2, 2000.

Contributed by Willem J. M. Levelt, August 29, 2001

A core operation in speech production is the preparation of words from a semantic base. The theory of lexical access reviewed in this article covers a sequence of processing stages beginning with the speaker's focusing on a target concept and ending with the initiation of articulation. The initial stages of preparation are concerned with lexical selection, which is zooming in on the appropriate lexical item in the mental lexicon. The following stages concern form encoding, i.e., retrieving a word's morphemic phonological codes, syllabifying the word, and accessing the corresponding articulatory gestures. The theory is based on chronometric measurements of spoken word production, obtained, for instance, in picture-naming tasks. The theory is largely computationally implemented. It provides a handle on the analysis of multiword utterance production as well as a guide to the analysis and design of neuroimaging studies of spoken utterance production.

The human ability to speak is universal. All normal children acquire the language of their environment at a very early age. Most start babbling at the age of 7 months, produce a few meaningful words around their first birthday, reach a 50-word vocabulary 6 months later, produce their first multiword utterances by the end of their second year of life, and begin expressing syntactic relations by means of prepositions, auxiliaries, inflections, and word order in the course of their third year. By the age of 5 or 6, the basic architecture of this natural skill is essentially in place. Although our ability to speak has since millennia been recognized as uniquely human, as species-specific, as the basis of our cultural evolution, and generally as a core aspect of the human condition (*homo loquens*), the systematic study of how we speak did not begin before the end of the 19th century. In 1900, Wilhelm Wundt (1) published his theory about how a sentence emerges in the speaker's mind, a theory entirely based on introspection. With their 1896 monograph, Meringer and Mayer (2) initiated an important empirical paradigm. They collected and analyzed a large corpus of spontaneously produced speech errors that they had carefully noted down. One of their findings was that word substitutions were either meaning-based [e.g., *Ihre* (your) for *meine* (mine)] or form-based [e.g., *Studien* (studies) for *Stunden* (hours)], suggesting a distinction between meaning- and form-based operations in word generation. It was only by the 1970s that this paradigm became fully exploited to construct theories of utterance generation (see ref. 3 for a review).

A core component of any such theory concerns lexical access. Although our speaking rate varies substantially, it is quite normal for a speaker to produce 2 to 4 words per second (4) and that is a surprising accomplishment. Apparently, we can access the appropriate words at this rate in our lexical memory, the "mental lexicon." Miller (5) estimated that the mental lexicon contains some 50–100,000 words in a normal literate adult person. The accessing system is, moreover, robust. On average, we err no more than once or twice every 1,000 words. We retrieve these words with their syntactic properties; these features play a crucial role in the incremental construction of the syntax of our utterance (4). Ultimately, each of these words must be given articulatory shape in the context of the larger utterance. That requires accessing a word's form properties, their "phonological codes" in memory. These are used by the speaker to compute or access the articulatory gestures for successive syllables. Syllables are major units of articulation; they easily come at a rate of 3–6 per second.

In the following I will outline a theory of lexical access that my research unit has been developing, building on detailed existing theories (reviewed in refs. 3 and 4), and in cooperation with many colleagues. The theory is one under development. It presently covers operations from the initial focusing of the speaker on a concept to be expressed "down" to the syllabification operations that precede the initiation of articulation. Roelofs (6, 7) provided the computational implementation of the theory as WEAVER++. Although the theory is inspired by speech error evidence, it is empirically largely based on reaction time data, in particular on speakers' word production latencies as measured in the laboratory. I will first sketch the architecture of the system, then discuss some aspects of its two major components: a system for lexical selection and a system for form encoding. I will then turn to some issues of repeated or successive lexical access as they occur in normal multiword utterance generation. Finally, I will discuss some applications of the theory in neuroimaging approaches to spoken word generation.

A Serial Two-System Architecture

Fig. 1 depicts the theory in a nutshell. To produce a content word, a speaker will first select the appropriate item from the mental lexicon. This is "lexical selection." Next, the selected item's articulatory shape will be prepared. This is "form encoding." Let us first consider lexical selection. Imagine the following experimental task: A subject is shown a picture of a horse and asked to name it. It depends on the subject's interpretation of the task what the response will be: "horse" is an obvious possibility, but it is not wrong to say "stallion" or "animal" (and some subjects do). The subject judges how much detail the experimenter would appreciate. More generally, a first step in preparing a content word is to focus on a concept whose expression will serve a particular communicative goal. This process is called "perspective taking" (4, 8). Some concepts are lexical concepts, those for which there is a lexical item in the mental lexicon. To initiate lexical selection, the speaker must focus on a lexical concept. I will denote lexical concepts in capitals: HORSE, STALLION, and ANIMAL. The theory assumes that during perspective taking there is coactivation of related concepts (Fig. 2). Each active lexical concept spreads activation to the corresponding lexical item in the speaker's mental lexicon. That item is called a "lemma." It is essentially the lexical item's syntactic description. For instance, the lemma for HORSE, *horse* (I will use italics to denote lemmas) specifies that it is a count noun and it has a variable diacritic for number (singular vs. plural). Its French equivalent (CHEVAL) would also be specified for gender (masculine). Although different lemmas (for instance HORSE and GOAT) may be syntactically identical, they are unique in their pointer to a phonological code (see below). The target lemma, i.e., the one activated by the focused concept, is selected under competition (6). The selection latency depends

Abbreviations: SOA, stimulus onset asynchrony; VT, viewing time; LRP, lateralized readiness potential; MEG, magnetic encephalography.

*E-mail: pim@mpi.nl.

www.pnas.org/cgi/doi/10.1073/pnas.231459498

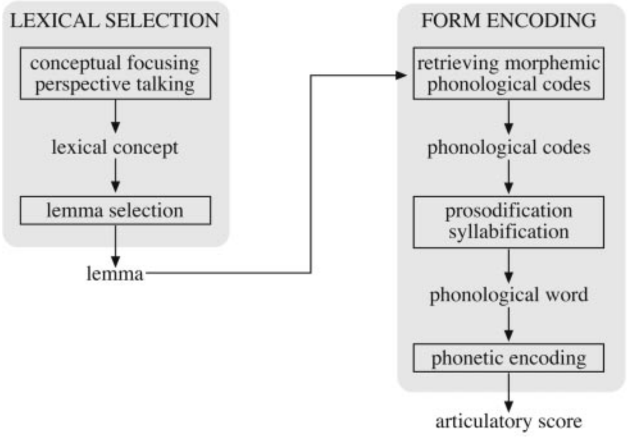

Fig. 1. Serial two-system architecture of the theory: two stages of lexical selection followed by three stages of form encoding.

on the amount of coactivation of other lemmas (such as *stallion* and *animal*). Lexical selection is complete as soon as the target lemma is selected.

This selection triggers the form encoding system (Fig. 3). Activation spreads from just the selected lemma to the phonological codes it points to; no other codes get coactivated (for instance, the codes of the coactivated lemmas *stallion* and *animal* remain silent). Many lemmas, for instance *horse* when marked for plural, have a multimorphemic code. A phonological code is retrieved for each of the morphemes, e.g., <horse> and <iz>, respectively. Phonological codes are "spelled out" as ordered sets of phonological segments, for instance/h, ɔ, r, s/and /ɪ,z/.[†] This forms the input to the operation of "prosodification," which is largely syllabification. The ordered segments are incrementally strung together to form legal syllables. In the example, a first syllable (σ) is created out of /h/, /ɔ/, and /r/:/hɔr/$_{\sigma}$ and then a second one out of /s/, /ɪ/ and /z/:/sɪz/$_{\sigma}$. This completes the syllabification of the phonological word (ω): /hɔr.sɪz/$_{\omega}$. Syllabification is context dependent: /hɔr/$_{\sigma}$ is a syllable of "horses" but is not one of "horse," whereas /hɔrs/$_{\sigma}$ is a syllable of "horse," but not of "horses." An item's syllabification is not stored in the mental lexicon but created on the fly, dependent on the current context. As syllables are incrementally composed, they are input to a final encoding step, phonetic encoding (Fig. 3). A core assumption of the theory is the existence of a "mental syllabary." This is a repository of highly practiced syllabic gestures (10). As syllabification proceeds, the corresponding syllabic patterns are selected from the syllabary for execution. Phonetic encoding also involves the smooth concatenation of retrieved syllabic routines. The string of syllabic gestural routines that corresponds to the target phonological word is called its "articulatory score." It is the output of form encoding and the final product of lexical access. The execution of successive articulatory scores by the speaker's laryngeal and supralaryngeal apparatus, articulation (as yet outside the theory, but see ref. 4), creates overt speech.

It is important to notice that the two systems involved in lexical access (Fig. 1) perform radically different functions. The function of lexical selection is to rapidly focus on a single lexical item, given the speaker's intentional state. This selection is subject to competition. The function of form encoding is to generate an articulatory score for just the selected item in its context. Competition is hardly an issue here. Initially, the two systems mature independently in the child's mind (11) and they involve different cerebral networks (12). The link between the systems is vulnerable. We all get occasionally into so-called "tip of the tongue" states, where the phonological code of a selected target item is temporarily partly or wholly unavailable, whereas syntactic, i.e., lemma information, is largely preserved (13). The rift between the systems is even magnified in certain anomic patients whose utterance production is normal, except that they block time and again on accessing the phonological codes of their target words (14). Finally, the serial two-system architecture predicts that lexical selection precedes form encoding. Supporting electrophysiological evidence will be discussed in the final section of this article.

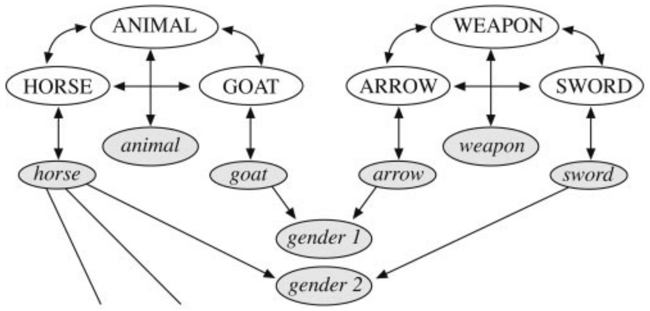

Fig. 2. Fragment of the WEAVER++ lexical selection network. (*Upper* stratum) Lexical concept nodes. (*Lower* stratum) Lemma and gender nodes.

Aspects of Lexical Selection

A Case of Perspective Taking. Much of the original work on perspective taking concerned the ways in which speakers conceptualize spatial states of affairs for their listeners (refs. 15 and 16; see ref. 17 for a comprehensive treatment of spatial perspective systems). If one asks subjects to describe the spatial pattern in Fig. 4*A*, which is put flat on the table in front of them, one gets two types of responses, depending on the perspective taken (18). Subjects taking the "intrinsic" perspective begin their description with something like "from the yellow dot you go straight to a green dot." The directional term is "straight." On successive moves, this is followed by "straight," "right," "straight," "right,"

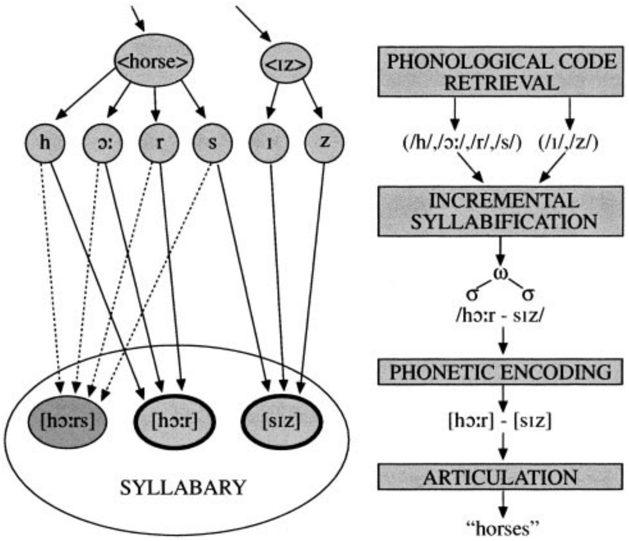

Fig. 3. Fragment of the WEAVER++ form encoding network (*Left*) with corresponding form-processing stages (*Right*). (*Upper* stratum) Nodes representing morphemic phonological codes and their phonemic "spellouts." (*Lower* stratum) Nodes representing syllabic articulatory scores.

[†] The plural morpheme code for English is factually more abstract. I refrain from discussing in detail its realization as /ɪz/ (see ref. 9).

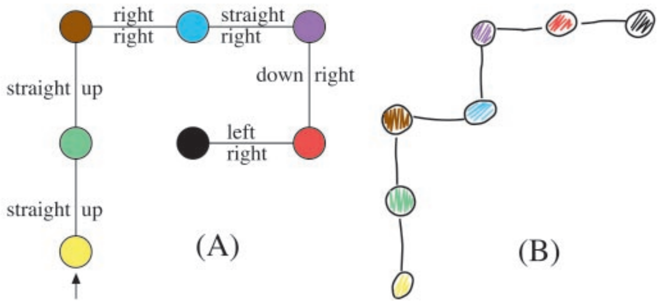

Fig. 4. (*A*) Experimental visual pattern to be described. Pattern is flat on the table in front of the subject. Description starts from the arrow. On the outside, directional terms used in "intrinsic" perspective. On the inside, terms used in "relative" perspective. (*B*) A typical subject's drawing of the same pattern from listening to an intrinsic description.

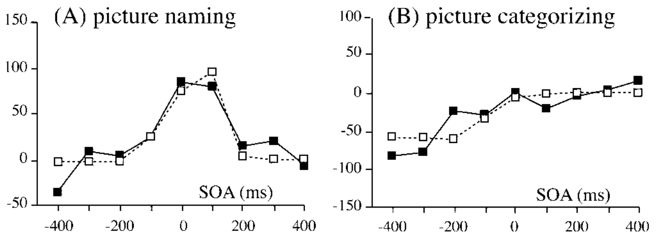

Fig. 5. Effects of a semantic distracter on picture-naming (*A*) and picture categorization latencies (*B*) for different SOAs of picture and visual distracter words. The semantic effect is the naming latency when the distracter is semantically related to the picture name minus the latency when the distracter word is unrelated to the picture name. Black squares represent data from Glaser and Düngelhoff (21). Open squares represent WEAVER++ simulations (6).

and "right." The subject makes a mental "body tour," relating each new direction to the direction of the previous move. The description is invariant over rotations of the pattern on the table; it is intrinsic to the pattern, given the path of the tour. However, the majority of subjects take the relative perspective. They interestingly begin their descriptions with a vertical dimension term "up" ("from yellow you go *up* to green"), followed by "up," "right," "right," "down," and "left." The vertical dimension terms emerge from the subjects making a "gaze tour." At each move they tell you where their gaze is going relative to their own oriented body position (16, 19). A relative description changes entirely when you rotate the pattern by 90, 180, or 270° on the table. Both descriptions are veridical, but they are almost entirely different. The final direction is even "right" in the intrinsic description and "left" in the relative description. Here the target concept RIGHT and the target concept LEFT denote the same state of affairs, given the perspectives in which they function. The formal properties of the two perspectives are drastically different. For instance, transitivity and converseness hold for the relative but not for the intrinsic system (18). Speakers typically do not tell their addressees what perspective they are using, which can lead to major confusion. When we asked subjects to listen to an intrinsic description of the pattern in Fig. 4*A* (without seeing it) and to draw it from the description, they typically drew the pattern in Fig. 4*B*. Here they forced a relative interpretation onto an intrinsic description (18).

Lemma Selection. Given a target concept, such as LEFT or HORSE, the subject will proceed to select the corresponding lemma. Let us assume that, in a picture-naming experiment, HORSE is the subject's target concept (Fig. 2). It spreads part of its activation to the lemma node *horse*, which is the one to be selected. HORSE also sends part of its activation to semantically related concept nodes, such as those for ANIMAL and GOAT. In turn, these spread part of their activation to their lemma nodes, *animal* and *goat*. Hence, the lemma *horse* is selected under competition (at least if "animal" and "goat" are permitted responses in the experiment). The equations in the computational model WEAVER++ that govern activation spreading and decay, as well as selection probabilities, are given in refs. 6 and 10. The basic notion is that during any minimal time unit, the probability of selecting the target lemma equals the ratio of its activation to the sum activation of all lemmas. This follows Luce's (20) choice rule and it predicts the expected selection latency. The prediction can be experimentally tested by manipulating lemma coactivation. The classical procedure is to present a printed distracter word in the picture to be named (6, 21) or to present an auditory, spoken distracter word (22) while the subject is naming the picture. Presenting the semantically related word "goat" as a distracter when HORSE is the target concept, the competing lemma *goat* is activated by the distracter (Fig. 2). In addition it receives activation through activation spreading from HORSE to GOAT to *goat*. If an unrelated distracter word is presented, such as "sword," that will activate the lemma *sword*, which will also compete with target lemma *horse*. However, *sword* does not receive extra activation through activation spreading from HORSE (as was the case for *goat*). Hence, the response "horse" will be slower when "goat" is the distracter than when "sword" is. This is called semantic interference. Another experimental manipulation is to vary the stimulus onset asynchrony (SOA) between distracter and picture; the distracter can be presented simultaneously with the picture, earlier or later. As Roelofs (6) has shown, the dynamics of activation spreading in WEAVER++ predict the amount of semantic interference for different SOAs. Fig. 5*A* presents the classical measurements by Glaser and Düngelhoff (21) with printed distracters along with the WEAVER++ simulation. Still another manipulation is to change the task from naming to categorization. The subject is instructed to take a different perspective, namely to say "animal" when seeing a horse or "weapon" when seeing a sword. Here WEAVER++ predicts facilitation by semantically related distracters for negative SOAs, and that fits the data obtained by Glaser and Düngelhoff (see Fig. 5*B*). Meanwhile, the body of evidence in support of the account by WEAVER++ of lexical selection is substantial (see ref. 10 for a review). Moreover, model and data provide estimates of the detailed time course of lexical access.

Lexical selection is not always "conceptually driven." In preparing the sentence, "I thought that Peter would be here," selection of the lemma *that* is triggered by a syntactic operation, not by focusing on a lexical concept (4).

After selection of the lemma, its syntactic properties are accessible for further processing. Grammatical encoding, the construction of phrases, clauses, and whole sentences, depends entirely on this information (4). A simple example case is the construction of an adjective-noun phrase, such as "big arrow." The syntax of lemmas *big* and *arrow* allow them to "unify" in that order—see ref. 23 for a computational model of unification. In many gender-marking languages, such as German or Dutch, the gender of the noun gets marked on the adjective during unification (German, gross***er*** Pfeil; Dutch, grot***e*** pijl). The gender of the adjective lemma is variable, the gender of a noun is fixed. To set the gender value of the adjective, the speaker must retrieve the gender of the noun. The faster this is done, the shorter the encoding latency of the adjective. It is experimentally possible to facilitate gender access (24–26), expressed doubts (27) notwithstanding. Recently, Vigliocco *et al.* (28) achieved this facilitation by means of the following paradigm. Dutch bilingual subjects produced phrases such as "grote pijl" (big arrow) or "kleine pijl" (small arrow) in response to an English stimulus word on the

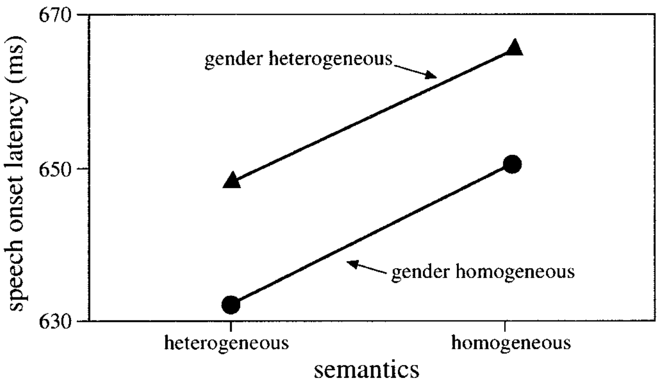

Fig. 6. Additive effects of semantic competition and gender facilitation in lexical selection. The target utterance, a Dutch adjective-noun phrase, is produced in response to an English probe word. The target nouns in an experimental block are either semantically homogeneous or heterogeneous. The syntactic gender of the target nouns in a block is either homogeneous (all gender 1 or all gender 2) or heterogeneous (both genders mixed). Data is from Vigliocco *et al.* (28).

screen (**ARROW** or **arrow**, respectively). Dutch has two genders, which are marked on the adjective as a schwa or as a null suffix ("grote"/xrɔ:t/ vs. "groot"/xrɔ:t/); they are not marked on the noun. In one condition of the experiment, target words of both genders were mixed 50/50; the gender was heterogeneous. In the other two conditions, all words had the same gender, either the one or the other; gender was homogeneous. The heterogeneous and homogeneous conditions involved the same words. Response latencies were a reliable 16 ms faster in the homogeneous than in the heterogeneous condition. Fig. 2 shows how this could have occurred. In Dutch, "pijl" (arrow) and "geit" (goat) have the same gender (gender 1). Repeated activation of that gender node will speed up its retrieval; this is called "gender priming." But mixing gender 1 items, such as "pijl" (arrow), with gender 2 items, such as "paard" (horse), annihilates the effect.

We could also use the translation paradigm to induce semantic competition (28). In a semantically homogeneous condition, the subjects translated items from the same semantic category, for instance all weapons, all animals, all plants, etc. In a semantically heterogeneous condition, the set of items was semantically mixed (a weapon, an animal, a plant, etc.). Again, the same items were involved in the two conditions, and the translation response was an adjective-noun combination. Response latencies were significantly longer (by 33 ms) in the homogeneous than in the heterogeneous condition, and this is what the WEAVER++ model predicts. Fig. 2 shows that there is activation spreading among same-category lexical concepts, i.e., among the animals and among the weapons but not between different-category items. This difference leads to smaller Luce (20) ratios in the homogeneous than in the heterogeneous conditions, and hence to longer selection latencies.

What would happen if we orthogonally varied semantic competition and gender priming in the same experiment? The theoretical option I explored is this: The speaker performs two successive operations, lemma selection and gender access [the latter is run only if the task requires it (see refs. 24 and 29)]. Semantic competition affects the first operation, gender priming the second one. This predicts additive effects if the two experimental manipulations are combined (30) and that is what I found (ref. 28; see Fig. 6). This interpretation of two successive operations was recently confirmed by Schriefers and Teruel (26). They showed, by means of picture word interference experiments, that the peak interference effect from an auditory semantic distracter occurs at earlier SOAs than the peak effect of a gender congruous distracter.

It would be premature to generalize these findings on gender access to each and every syntactic lemma property, such as number, person, or tense. The theory (10) leaves this open. In particular, it does not make the claim, neither for gender nor for any other syntactic feature, that it is selected under competition; we propose no Luce (20) ratio rule for the selection of feature values. The ways speakers access the syntax of item is still the least explored aspect of lexical selection.

Aspects of Form Encoding

Accessing the Phonological Code. A first step in crossing the rift from lexical selection to form encoding is to access the phonological code(s) of the selected item (Figs. 1 and 3). W.J.M.L. *et al.* (31) formulated as a core assumption of the theory that only a selected lemma spreads its activation across the rift to its phonological code. This hypothesis was based on the results of a picture-naming experiment, which showed that upon seeing a picture of an object to be named (for instance a horse), the subject activated the corresponding lemma as well as lemmas of semantically related concepts (such as *goat*). The subject also activated the phonological code of the target word (<horse>), but never to any extent the code of an active semantic alternative (such as <goat>). This assumption has been repeatedly challenged, and ultimately one exception turned up (32, 33). If you name an object with two synonym names, for instance a couch which can also be called "sofa," both phonological codes can get simultaneously activated. My theoretical concession has been that double lemma selection is a possibility in case of synonymy (10). This hypothesis is consonant with speech error data. Synonym blends, such as "clear" for "close" and "near", are not uncommon, whereas semantic alternatives (such as "horse" and "goat") hardly ever blend. Thus far, no other exceptions have been found.

A further important aspect of accessing the phonological code is that it is word frequency (WF)-dependent. Accessing the code is slower for low-frequency than for high-frequency words (34). The frequency of usage of a word is highly correlated with the age at which it has been acquired by the child (AoA), and these parameters are hard to disentangle. Hence, WF dependency can in part or in full be AoA dependency (10).

Accessing the phonological code can be manipulated in a picture-naming experiment by presenting an auditory distracter that is or is not phonologically related to the target name. For instance, if the picture is one of a book, the spoken distracter word can be "boor" (begin-related to "book"), "look" (end-related to "book",) or "fan" (unrelated). Related distracters produced shorter naming latencies than unrelated ones (22, 35). This phonological facilitation effect was further explored by Meyers and Schriefers (36) for both begin- and end-related distracters. Fig. 7 presents a subset of these data as well as Roelofs' (7) WEAVER++ simulations thereof. It is not essential that the distracter is a word itself. Related nonwords also produce facilitation. For instance, naming a hammer is facilitated by presenting its last syllable ("mer") as a distracter (7). It is further irrelevant whether the phonologically related distracter is itself a whole syllable of the target word (37, 38). This finding supports the notion that for the tested languages (Dutch, German, and English), syllable structure is not represented in the phonological code.

Taken together, these data support the theoretical assumption, implemented in WEAVER++, that the phonological code is retrieved as a whole. Priming any of its parts facilitates retrieval of the entire code. However, it is likely that retrieving the codes of multimorphemic words is performed incrementally, code after code. Roelofs (39) showed this to be the case for com-

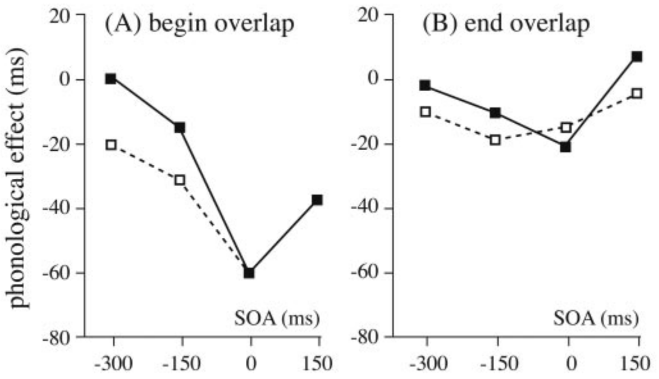

Fig. 7. Effects of a phonological distracter on picture-naming latencies, for different SOAs between picture onset and onset of spoken distracter. All picture names and distracter words were disyllabic. (*A*) Distracter and target words share their first syllables. (*B*) Distracter and target words share their last syllables. The phonological effect is the naming latency when the distracter is phonologically related to the target name minus the naming latency when the distracter is unrelated. Black squares represent data from Meyer and Schriefers (36). Open squares represent WEAVER++ simulations (7).

pounds. In fact, the morphology of a target word is a crucial factor in phonological encoding, as will be elaborated below.

Incremental Syllabification and Morphological Frames. Syllabification may run over morpheme boundaries (as in "horses"/hɔr.sɪz/$_{\omega}$) and even over lexical boundaries (as in "he owns it"/oʊn.sɪt/$_{\omega}$). In English and many other languages, the syllabification of a word is context-dependent. Syllabification boundaries are determined by various syntactic and morphological conditions (4). The central theoretical claim is that syllabification is computed incrementally over the spelled-out phonological codes. The substantial experimental evidence proceeds from a paradigm introduced by Meyer (40, 41), "implicit priming" (Table 1). The subject learns a small set of semantically related probe-target pairs, for instance "single-loner," "place-local," and "fruit-lotus" (Table 1, Set 1; the target words are displayed only). Then the probe words are repeatedly presented on the screen, in random order. Each time the subject responds by producing the corresponding target word. Speech onset latencies are measured. Next, a second set is learned (Table 1, Set 2; target words "beacon," "beadle," and "beaker"), and the response procedure is repeated for this set. Finally, the procedure is run on a third set (Table 1, Set 3). Each of these sets is homogeneous in that the target words begin with the same syllable ("lo," "bea," and "ma," respectively). The speaker implicitly knows the first syllable of the target word. To check whether this facilitates the encoding of the first syllable of the target word, the same words are arranged into three different sets (sets 4, 5, and 6 in Table 1). Target words in these sets do not share their first syllables. Will onset latencies be slower here? They are indeed, by some 30 ms. The crucial test for incrementality of encoding is to use the same paradigm on sets of target words that share a noninitial stretch. This is the case for sets 7–9 in Table 1; the target words share their last syllables ("to," "ble," and "va," respectively), but their initial syllables are different. If subjects can use their implicit knowledge of the second syllable to speed up encoding, they should be faster on sets 7–9 than on sets 10–12, which do not share the final syllable. But this is not what happens. There is not even a tendency for response latencies to be shorter for sets 7–9. The general finding turned out to be that there is never any implicit priming for words which share a noninitial stretch, not even for monosyllabic rhyme words such as "dock," "lock," and "rock." Syllabification is strictly incremental; you can only prepare a word from its start. There is measurable implicit priming if a target set shares only its first segment {"pen," "pot," "pal"}, more if it shares a first syllable, and even more if it shares two syllables out of three, etc. But priming less than a segment is impossible. If the target set is {"ben," "pot," "pal"}, the initial segments vary only in one phonological feature, voicing. That is enough to entirely annihilate implicit priming. You can prepare the initial segment of a word but not a subset of its phonological features, as Roelofs (42) has shown.

Table 1. Implicit priming of first (sets 1–3) and second (sets 7–9) syllable

Set	S1	S2	S3		S7	S8	S9
S4	*loner*	*beacon*	*major*	S10	*salto*	*feable*	*lava*
S5	*local*	*beadle*	*maker*	S11	*veto*	*stable*	*nova*
S6	*lotus*	*beaker*	*maple*	S12	*photo*	*rubble*	*diva*

Let us return to the encoding of multimorphemic words, such as "horses." Previously, I referred to evidence that the phonological codes of the morphemes of a word are successively retrieved. In fact, the morphological structure of the target word plays a crucial role in phonological encoding. Janssen *et al.* (43) demonstrated this idea for Dutch inflectional morphology. Again, using the implicit priming paradigm, they showed that priming by a shared initial syllable is stronger when the target words share their inflectional affix structure than when they do not. The experiments support the notion proposed in ref. 4 that, right after lemma selection, a morphological target frame is constructed (for instance "stem+affix+affix") into which the incrementally retrieved morphophonological codes are inserted. The process of constructing that frame can be implicitly primed. One function of constructing such a frame is to delineate the domains of syllabification. For English, a stem+affix frame typically indicates a single domain of syllabification, i.e., the domain of a phonological word, as is the case for /hɔr.sɪz/$_{\omega}$. But the stem+stem frame of compounds (such as "popart") induces two successive domains of syllabification. Each stem gets realized as a single phonological word; the speaker first syllabifies the first stem (/pɔp/$_{\omega}$) and then the second one (/a:^{r}t/$_{\omega}$). This syllabification has as a consequence that the middle /p/ of "popart" is not realized as the first segment of the second syllable (as was the case for /s/ in "horses"). That we know because that /p/ does not get aspirated, as a syllable-initial /p/ does.

Languages differ substantially in morphology and syllabification. The crosslinguistic study of incremental phonological encoding is only beginning. Chen *et al.* (44) used implicit priming to study incremental syllabification in Mandarin Chinese. Each syllable has one of four tones in Chinese. Word initial-syllable priming turned out to be possible, independent of tone. But tone priming alone was ineffective. Still, shared tone does contribute. The strongest priming was obtained when the target words in a set shared their first syllable and its tone. Hence, tone priming is conditional on shared phonemic or syllabic content. It should not be surprising if form encoding differs substantially between Chinese and languages such as English or Dutch. In Mandarin Chinese, syllabification is not context-dependent, and the number of different syllables is exceedingly small (some 400, not counting tones). That would make it efficient for speakers to store syllable boundaries in the form lexicon and dispense with computing these boundaries on the fly during syllabification.

Phonetic Encoding. Like Mandarin Chinese, many languages have no more than a few hundred different syllables. Assume you talk, from your 2nd to your 21st birthday, on average 30 min a day. If you speak at an average rate of 4 syllables per second, you will have produced 5.10^7 syllable tokens at reaching adulthood. If your language counts 500 different syllables, each of them will, on average, have been produced some 10^5 times. This is a classical case of "overlearning." Articulatory syllables are among the most exercised motor patterns we produce. Such overlearned motor actions typically get stored in the premotor cortex (45). I have called this hypothetical repository the "mental syllabary" (46, 47). But what if a language has many more different syllables? English and Dutch, for instance, have far more than 10,000. Statistics show that speakers of these languages do 80% of their talking with no more than 500 different high-frequency syllables (10). Hence, these are hardly less overlearned as articulatory motor actions, and thus likely members of a speaker's syllabary. Although we rarely produce a syllable that we never produced before, the low-frequency tail of the distribution may contain syllables which never became stored as motor patterns. The articulation of such items should be prepared on the fly from the syllabified phonological word (ω). Here I will limit to our modeling of syllabary access (7, 10). It is exemplified in Fig. 3. As soon as a phonological code is retrieved (for instance the code /hɔrs/), its segments activate all syllabic gestures in the syllabary in which they partake. As an example, the spelled-out segment /s/ spreads activation to the articulatory syllables [hɔrs], [sɪz], and many others. There will, at any one moment, be just one target syllable, given the state of incremental phonological syllabification. Selection of that syllabic gesture occurs under competition, following the same selection mechanism as we saw above for lemma selection. But different from what was the case for lemma selection, the chronometric experimental support for the mechanism of syllable selection is still fragmentary.

Multiple Lexical Access

Initiating Multiword Phrases. Given the serial two-system architecture of lexical access, lexical selection followed by form encoding, one can now ask how access is organized in simple multiple-word utterances. Above we saw that it is possible to prime the gender of the noun in the production of a simple Dutch phrases such as "grote pijl"—"big arrow" (26, 28). Here, the gender of the noun must be marked on the adjective. The speaker cannot complete the form encoding of the adjective without having accessed the lemma of the noun, in particular its gender feature. Hence, in this case, form encoding of item 1 depends on lemma access to item 2, and that has a measurable effect on the initiation of articulation. But even without such dependency, as is the case in English, the utterance is not initiated before the head noun has been selected. For instance, Costa and Caramazza (48) had subjects produce phrases such as "the red car" in response to a picture. Simultaneously presented (visual) semantic distracter words (such as "truck") delayed response initiation as compared with unrelated distracters. Meyer (49) had obtained the same result for more complicated locative utterances, such as the Dutch equivalent of "the baby is next to the dog" (in response to a two-object picture as in Fig. 8). When subjects were given an auditory semantic distracter related to "dog" (for instance "cow"), their response was slower than when the distracter was unrelated (for instance "pipe"). However, this effect is not always obtained in utterances with multiple noun phrases. Here the planning window for lexical selection is variable and strategic (50).

What about the planning window of form encoding? If item 2 is selected before utterance initiation, is it also form encoded? Meyer (49) showed that not to be the case for her locative sentences. In the example, a phonological distracter such as "doll" related to item 2 ("dog") did not facilitate speech onset for the utterance as a whole. Initially, the same was found for adjective-noun phrases such as "(the) red car": phonologically priming the noun did not facilitate onset latency (24, 51). More recently, however, Costa and Caramazza (48) did obtain such an effect, which confirms earlier findings by Wheeldon and Lahiri (52), using a different paradigm.

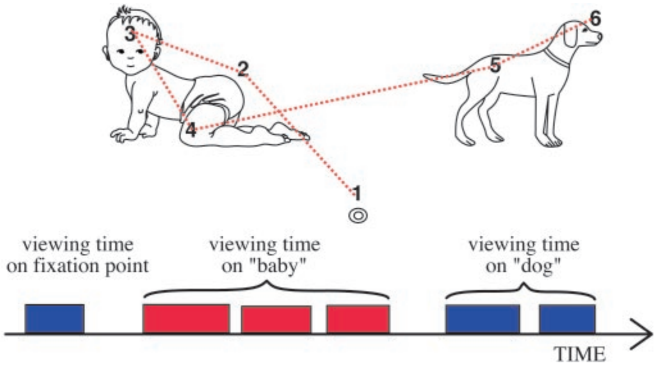

Fig. 8. Visual scan during a scene description. As soon as the two-object picture appears, the subject makes a saccade from the fixation point to the left object. After some scanning of that object, the gaze shifts to the right one. VT is the total duration of fixating on an object. The utterance produced here is "baby and dog."

Coordinating Lexical Selection and Form Encoding During Utterance Generation. Utterance onset measurements cannot reveal the coordination of lexical selection and form encoding after speech onset. Meyer *et al.* (53) introduced an eye-scanning paradigm to study utterance encoding in scene-description tasks. Fig. 8 exemplifies a typical experimental trial. The subject looks at a fixation point until a picture appears. The instruction is to describe the scene with an utterance such as "baby and dog." Meanwhile, the subject's eye movements are being monitored by an eye-tracking system. Fig. 8 displays an actual eye scan obtained during such a trial. The experimental output consists of the scanning pattern and the speech recording, both time-locked to picture onset. One important variable is "viewing time" (VT). VT is the duration of looking at an object (for instance the baby in Fig. 8). It begins with the first fixation on the object and ends with the saccade out of the object (toward the dog). The characteristic scanning pattern is that subjects view the objects, whether two or more, in the order of mention. That reflects their pattern of attending to the objects. The initial hypothesis by Meyer *et al.* was that subjects would visually attend to an object just long enough to select the appropriate lexical item. Form encoding could then follow while visual attention turns to the second object. But the experiments turned out otherwise. VT for the left object covaried with the difficulty of form encoding. Remember that the speed of accessing phonological code of a word is word frequency-dependent (34). When Meyer *et al.* compared scenes where the left object had a high-frequency name as opposed to a low-frequency name (controlling for visual recognizability), they obtained a corresponding difference in VTs of 28 ms. Similarly, priming the phonological encoding of the left object's name ("baby" in Fig. 8), by presenting the subject with a phonological distracter word (for instance "lobby"), shortened VTs on average by a reliable 36 ms (54). VTs were also longer for objects named with a full noun than for objects named with a pronoun (such as "it"; ref. 55). Taken together, these data suggest that the speaker keeps visually attending to the target object until phonological encoding, including syllabification, is complete (50). The functional significance of this strategy is a matter of dispute. Speakers may construct successive "attentional islands" to minimize interference between subsequent same-kind operations (lexical selection, phonological encoding). Such interference can cause speech errors. On the other hand, visual attention should shift early enough to keep the utterance fluent. The balance is a subtle one, and speakers can fail both ways.

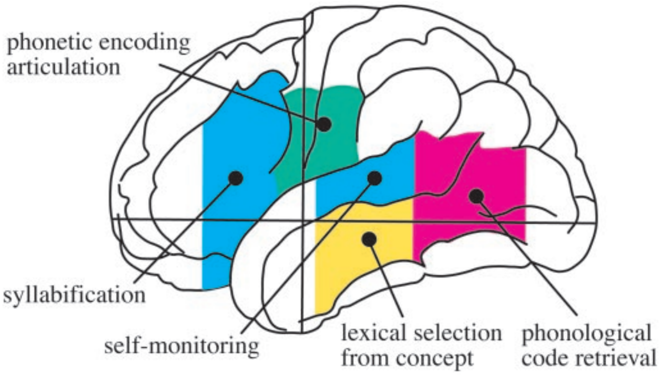

Fig. 9. A metaanalysis of 58 neuroimaging studies of word production. Colors denote regions whose activations correspond to theoretical processing stages as indicated. Contributions of insula, subcortical regions, and cerebellum are not shown. [Reproduced with permission from ref. 64 (Copyright 2000, MIT Press, Cambridge, MA).]

Some Neurological Perspectives

The theory has been a guide to chronometric studies of cerebral activation in word production. It has also contributed to metaanalyses of localizational positron-emission tomography (PET) and functional MRI (fMRI) studies of word production.

Time Course Studies. The theory is largely based on chronometric studies of lexical access. Combined with Roelofs' computational modeling, the theory provided initial estimates of the time course of the various operations involved in lexical selection and form encoding (56). However, measurements of speech onset latency have limited potential in this respect. The reaction time is the cumulative effect of no less than five successive operations (Fig. 1). How to distinguish these operations in the temporal domain? In some cases, SOA studies provide the relevant data. For instance, the peak effect of a semantic distracter or probe during picture naming always precedes the peak effect of a phonological one (31, 35). The difference provides an estimate of the time lapse between the successive operations of lemma selection and phonological code retrieval. For most pairs of stages, however, such distracter studies are impossible.

Here the measurement of evoked electrical and magnetic brain responses has come to the rescue. Van Turennout *et al.* (57) were the first to measure lateralized readiness potentials (LRPs) to distinguish between stages of lexical access. This measure is based on the electrophysiological brain potential that precedes any voluntary hand movement. The LRP is measured time-locked to the stimulus, for instance a picture presented to the subject. The subject provides a push-button response to an aspect of the picture, indicating for instance whether it is animate or inanimate, whether its name begins with /b/or/v/or, in experiments on Dutch and German word production, whether the name has syntactic gender 1 or gender 2. Crucial for the experiments is that, under specific conditions, an LRP can be measured not only preceding an overt response, i.e., in the "go" condition, but also in a "no-go" condition. For instance, when the subject is instructed to respond to the gender of the word, but only in case the target word begins with /b/, not in case it begins with/z/, the latter condition still induces an LRP, which reveals the subject's preparation to give a gender response even in the "no-go" condition. The phonological code comes too late to prevent preparing a gender response. No such no-go LRP appears if the subject is instructed to respond to the initial letter of the word, /b/or/z/, but only in case the target word is of gender 1. According to the theory, the subject retrieves the gender of the word before its phonological code. The gender information is timely enough to prevent preparation of a phonological response. By using the LRP method, rather precise estimates can be made of the time course of semantic, syntactic, and phonological activations. For instance, van Turennout *et al.* (58) measured a 40-ms delay between accessing the gender of a word and accessing its first phoneme. Schmitt *et al.* (59, 60) added a further electrophysiological response to the time course measurement of word production, the N200 potential. It measures response inhibition, which is a central aspect of the just-mentioned experimental paradigm. One of their important findings was that conceptual focusing precedes gender access by some 80 ms.

Magnetic encephalography (MEG) has also been used to answer time course questions raised by the theory. In addition, whole-head MEG provides useful information about brain regions that are involved in successive stages of word production. W.J.M.L. *et al.* (56) used a picture-naming task to explore the chronometry derived from the theory in relation to the spatio-temporal distribution of brain activation from picture onset to onset of articulation. A core finding was that during the (estimated) time window of phonological encoding (275–400 ms after picture onset) peak activity was observed in a cluster of dipole sources in classical Wernicke's area (a left posterior temporal region), which is known to be a repository of word forms, i.e., phonological codes. In a recent MEG study, Maess *et al.* (61) specifically tested the operation of lemma selection. From earlier picture-naming studies, selection was estimated to take place around 200 ms after picture onset. They used a picture version of the translation paradigm discussed previously (29). In one condition, subjects named pictures in blocks from a single semantic category, for instance just vehicles or just animals (the "homogeneous" condition). In another condition, they named the same pictures in blocks consisting of different semantic categories, for instance containing a vehicle, an animal, etc. (the "heterogeneous" condition). Subjects were slower in the homogeneous than in the heterogeneous blocks and, as discussed previously, this is the result of competition in lemma selection. A principal components analysis (PCA) over the time courses of dipole activation (251 dipoles were "projected" 15 mm below the surface of the dura) produced one factor that significantly distinguished between the activation patterns obtained in the homogeneous and heterogeneous conditions. The difference appeared in a time window of 150–225 ms post picture onset. This finding confirms and further specifies the earlier chronometric estimates for this stage of processing[‡]. The relevant PCA factor reflects a region in the left lateral temporal lobe. This is in excellent agreement with metaanalysis findings (see below).

Metaanalyses. Positron-emission tomography (PET) and functional MRI (fMRI) studies of spoken word production have used a variety of experimental tasks, such as picture-naming and word-reading, and a wide range of brain regions have been reported to be involved in the execution of these tasks. Can this cerebral network be transparently related to the processing components of the present theory? Indefrey and W.J.M.L. (62, 63) performed metaanalyses of the published neuroimaging experiments on word production, using the processing theory to make the relevant contrasts. This approach required a detailed analysis of the experimental tasks used. For instance, in a picture-naming task there is a task-specific "lead in" process: visual object recognition. After object recognition, all "core" processes of lexical preparation (Fig. 1) follow suit, from perspective taking and lemma selection all the way to phonetic

[‡] The conditions also contrasted at the much later time window of 450–475 ms, shortly before speech onset. This result was adduced to semantic self-monitoring, not discussed in the present article, but see Fig. 9.

encoding. Another much-used task, verb generation (the subject hears a noun such as "apple" and produces a verb denoting a use for it, e.g., "eat"), has a different lead-in: auditory word recognition and imagining a use of the object. From there again all core processes of word production are run like in picture-naming. Compare these to a task such as word-reading. Here the task-specific lead-in process is word recognition, mapping the visual word onto the corresponding phonological code. After retrieval of that code, the remaining core processes, syllabification and phonetic encoding are run before articulation can be initiated. Note that there is no conceptually driven lemma selection here. By contrasting all imaging data from tasks that do involve lemma selection to tasks that do not (and by using a binomial statistical "filter"), one region was found that systematically distinguished between them: the mid-section of the left middle temporal gyrus. This finding is in fine agreement with the above-mentioned MEG study of lemma selection (61), which showed the left lateral temporal lobe to be involved with that operation.

The task analyses allowed us to construct other contrasts as well. For instance, we compared brain activations for tasks involving phonological code retrieval (such as picture-naming and word-reading) to tasks that do not (reading a nonword such as "flimtis"). This contrast showed the specific involvement of Wernicke's area in phonological code retrieval in agreement with the above-mentioned MEG findings (56). Fig. 9 summarizes the main findings of the original metaanalysis (62, 63) over 58 published data sets. A recent extension to 82 data sets confirms and refines these findings.

Prospects. These post hoc analyses display a satisfying transparency between the extensive network involved in the production of spoken words and the theory of lexical access reviewed in this article. Clearly, the theory has the potential to guide the construction of neuroimaging studies of word production which are substantially more specific than what has hitherto been done. In particular, it invites a directed approach to issues such as lemma selection, accessing syntactic features, morphological encoding, and syllabification, which have rarely been the subject of positron-emission tomography (PET) or functional MRI (fMRI) research.

I thank Antje Meyer and Ardi Roelofs, whose contributions to the theory are invaluable, for helpful comments.

1. Wundt, W. (1900) *Die Sprache* (Engelmann, Leipzig, Germany).
2. Meringer, R. & Mayer, K. (1896) *Versprechen und Verlesen* (Goschen, Stuttgart).
3. Levelt, W. J. M. (1999) *Trends Cogn. Sci.* **3,** 223–232.
4. Levelt, W. J. M. (1989) *Speaking: From Intention to Articulation* (MIT Press, Cambridge, MA).
5. Miller, G. A. (1991) *The Science of Words* (Scientific American, New York).
6. Roelofs, A. (1992) *Cognition* **42,** 107–142.
7. Roelofs, A. (1997) *Cognition* **64,** 249–284.
8. Clark, E. (1997) *Cognition* **64,** 1–37.
9. Halle, M. (1978) in *Linguistic Theory and Psychological Reality*, eds. Halle, M., Bresnan, J. & Miller, G. A. (MIT Press, Cambridge, MA).
10. Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999) *Behav. Brain Sci.* **22,** 1–38.
11. Levelt, W. J. M. (1998) *J. Psycholinguist. Res.* **27,** 167–180.
12. Levelt, W. J. M. (2001) in *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, ed. Dupoux, E. (MIT Press, Cambridge, MA).
13. Vigliocco, G., Antonioni, T. & Garrett, M. F. (1997) *Psychol. Sci.* **8,** 314–317.
14. Badecker, W., Miozzo, M. & Zanuttini, R. (1995) *Cognition* **57,** 193–216.
15. Miller, G. A. & Johnson-Laird, P. N. (1976) *Language and Perception* (MIT Press, Cambridge, MA).
16. Levelt, W. J. M. (1982) in *Speech, Place, and Action: Studies in Deixis and Related Topics*, eds. Jarvella, R. J. & Klein, W. (Wiley, New York).
17. Levinson, S. C. *Space in Language and Cognition: Explorations in Cognitive Diversity* (Cambridge Univ. Press, Cambridge, U.K.), in press.
18. Levelt, W. J. M. (1996) in *Language and Space*, eds. Bloom, P., Peterson, M. A. & Nadel, L. (MIT Press, Cambridge, MA).
19. Shepard, R. & Hurwitz, S. (1982) *Cognition* **18,** 161–193.
20. Luce, R. D. (1959) *Individual Choice Behavior* (Wiley, London).
21. Glaser, W. R. & Düngelhoff, F.-J. (1984) *J. Exp. Psychol. Hum. Percept. Perform.* **10,** 640–654.
22. Schriefers, H., Meyer, A. S. & Levelt, W. J. M. (1990) *J. Mem. Lang.* **29,** 86–102.
23. Kempen, G. (2001) in *The Syntax–Semantics Interface. Linguistic Structures and Processes*, eds. Härtl, H., Olsen, S. & Tappe, H. (De Gruyter, Berlin).
24. Schriefers, H. (1993) *J. Exp. Psychol. Learn. Mem. Cognit.* **19,** 841–850.
25. La Hey, W., Mak, P., Sander, J. & Willeboordse, E. (1998) *Psychol. Res.* **61,** 209–219.
26. Schriefers, H. & Teruel, E. (2000) *J. Exp. Psychol. Learn. Mem. Cognit.* **26,** 1368–1377.
27. Costa, A., Sebastián-Gallés, N., Miozzo, M. & Caramazza, A. (1999) *Lang. Cogn. Processes* **14,** 381–391.
28. Vigliocco, G., Lauer, M., Damian, M. F. & Levelt, W. J. M. *J. Exp. Psychol. Learn. Mem. Cognit.*, in press.
29. Damian, M. F., Vigliocco, G. & Levelt, W. J. M. (2001) *Cognition* **81,** B77–B86.
30. Sternberg, S. (1969) *Acta Psychologica* **30,** 276–315.
31. Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, Th. & Havinga, J. (1991) *J. Psychol. Rev.* **98,** 615–618.
32. Peterson, R. R. & Savoy, P. (1998) *J. Exp. Psychol. Learn. Mem. Cognit.* **24,** 539–557.
33. Jescheniak, J. D. & Schriefers, H. (1998) *J. Exp. Psychol. Learn. Mem. Cognit.* **24,** 1256–1274.
34. Jescheniak, J. D. & Levelt, W. J. M. (1994) *J. Exp. Psychol. Learn. Mem. Cognit.* **20,** 824–843.
35. Roelofs, A. *Quart. J. Exp. Psychol.,* in press.
36. Meyers, A. S. & Schriefers, H. (1991) *J. Exp. Psychol. Learn. Mem. Cognit.* **17,** 1146–1160.
37. Baumann, M. (1995) *The Production of Syllables in Connected Speech* (MPI, Nijmegen, The Netherlands).
38. Schiller, N. O. (2000) *J. Exp. Psychol. Learn. Mem. Cognit.* **26,** 512–528.
39. Roelofs, A. (1996) *J. Mem. Lang.* **35,** 854–876.
40. Meyer, A. S. (1990) *J. Mem. Lang.* **29,** 524–545.
41. Meyer, A. S. (1991) *J. Mem. Lang.* **30,** 69–89.
42. Roelofs, A. (1999) *Lang. Cogn. Processes* **14,** 173–200.
43. Janssen, D. P., Roelofs, A. & Levelt, W. J. M. *Lang. Cogn. Processes,* in press.
44. Chen, J.-Y., Chen, T.-M. & Dell, G. *J. Mem. Lang.,* in press.
45. Rizzolatti, G. & Gentilucci, M. (1988) in *Neurobiology of Motor Cortex*, eds. Rakic, P. & Singer, W. (Wiley, New York).
46. Levelt, W. J. M. (1992) *Cognition* **42,** 1–22.
47. Levelt, W. J. M. & Wheeldon, L. (1994) *Cognition* **50,** 239–269.
48. Costa, A. & Caramazza, A. *J. Mem. Lang.,* in press.
49. Meyer, A. S. (1996) *J. Mem. Lang.* **35,** 477–496.
50. Levelt, W. J. M. & Meyer, A. S. (2000) *Eur. J. Cogn. Psychol.* **12,** 433–452.
51. Schriefers, H. & Teruel, E. (1999) *Eur. J. Cogn. Psychol.* **11,** 17–50.
52. Wheeldon, L. & Lahiri, A. (1997) *J. Mem. Lang.* **37,** 356–381.
53. Meyer, A. S., Sleiderink, A. M. & Levelt, W. J. M. (1998) *Cognition* **66,** B25–B33.
54. Meyer, A. S. & Van der Meulen, F. (2000) *Psychon. Bull. Rev.* **7,** 314–319.
55. Van der Meulen, F., Meyer, A. S. & Levelt, W. J. M. (2001) *Mem. Cogn.* **29,** 512–521.
56. Levelt, W. J. M., Praamstra, P., Meyer, A. S., Helenius, P. & Salmelin, R. (1998) *J. Cogn. Neurosci.* **88,** 553–567.
57. Van Turennout, M., Hagoort, P. & Brown, C. M. (1997) *J. Exp. Psychol. Learn. Mem. Cognit.* **23,** 787–806.
58. Van Turennout, M., Hagoort, P. & Brown, C. M. (1998) *Science* **280,** 572–574.
59. Schmitt, B. M., Münte, T. F. & Kutas, M. (2000) *J. Mem. Lang.* **37,** 473–484.
60. Schmitt, B. M., Schiltz, K., Zaake, W., Kutas, M. & Münte, T. F. (2001) *J. Cogn. Neurosci.* **13,** 510–522.
61. Maess, B., Friederici, A. D., Damian, M., Meyer, A. S. & Levelt, W. J. M. (2001) *J. Cogn. Neurosci.,* in press.
62. Indefrey, P. & Levelt, W. J. M. (2000) in *The New Cognitive Sciences*, ed. Gazzaniga, M. (MIT Press, Cambridge, MA).
63. Levelt, W. J. M. & Indefrey, P. (2000) in *Image, Language, Brain*, eds. Marantz, A., Miyashita, Y. & O'Neil, W. (MIT Press, Cambridge, MA).