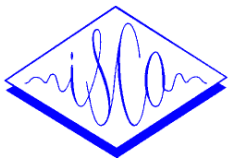

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ON THE SKILL OF SPEAKING: HOW DO WE ACCESS WORDS?

Willem J.M. Levelt

Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

ABSTRACT

Central to the skill of speaking is our ability to select words that appropriately express our intentions, to retrieve their syntactic and phonological properties and to compute the ultimate articulatory shape of these words in the context of the utterance as a whole. The generation of words in speech involves a number of processing stages. This paper discusses a stage of *conceptual preparation*, which leads to the activation of a lexical concept; a stage of *lexical selection,* which leads to the retrieval of an appropriate word or *lemma* from the mental lexicon; a stage of *phonological encoding* in which phonological words, consisting of phonological syllables are created; a stage of *phonetic encoding*, which produces a string of syllabic gestural scores, which can eventually be executed during the final stage of *articulation*.

I. INTRODUCTION

Any theory of how we produce words when we speak should be embedded in an account of how we speak at all. Normal adults are skilled speakers, who may be spending several hours a day in conversation with others and in talking to themselves. This skill allows us to generate utterances that reveal some intention, emotion, attitude, feeling, plan or whatever to our interlocutors. It also allows us to explicate our own internal world to ourselves in a medium that is surprisingly similar to the verbal medium of exchange with others; although the structure of internal speech is not well understood, it is a phonetic representation of some sort [1][2]. Before going into issues of lexical access proper, then, I must first embed them in this larger framework. As any complex skill, speaking involves the interaction of several processing components. Here is a bird's eye view of this system:

A first component is concerned with the speaker's conceptual preparation for speech. Speaking is a form of goal directed behavior. The speaker conceives of some communicative intention, and the goal is usually that the interlocutor will recognize it. In order to achieve this, speakers must select and order information whose expression will directly or indirectly reveal their intentions. This is called *macro-planning*. When we formulate a request, relate an event we observed, explain the game of chess or describe the layout of our home, we will elaborate our communicative intention in a sequence of goals and subgoals, selecting information whose expression will realize each subgoal, and in a communicatively effective order. The information to be expressed should also be given some kind of propositional format. This is called *micro-planning*. For instance, when we express spatial information, we must relate objects in the spatial image as in a proposition: *the cat is near the table* or *the table is near the cat*; there are canonical or preferred ways of doing this; we call this *perspective taking*. The output of conceptual generation consists of so-called *messages*, propositional structures that have lexical concepts as terminal elements. A lexical concept is one for which we have a word in the mental lexicon.

Whatever the message is going to be, it has to be cast in linguistic form. This is the task of the next component, the *formulator*. In fact, it consists of two subcomponents that perform quite distinct operations. The first operation is *grammatical encoding*. It consists of selecting the appropriate words from the lexicon, given the message, and creating morpho-syntactic order. Each retrieved word carries its own syntactic requirements. A verb "wants" to be in a verb phrase environment, a noun "wants" to be part of a noun phrase, etc. All these syntactic requirements of the different words retrieved should eventually be satisfied. Grammatical encoding is somewhat like solving a set of simultaneous equations. Not quite simultaneous, though, because words are not simultaneously retrieved. The order in which they get retrieved turns out to be a major determinant of the resulting syntactic structure [3].

The second operation is to generate a phonological, and ultimately articulatory-phonetic shape for each word and for the utterance as a whole. This is *phonological* and *phonetic encoding*. The speaker generates what is known as the prosodic hierarchy: syllables, phonological words and phrases, and intonational phrases. Phenomenologically, the output of this component, the phonetic plan, presents itself to us as *internal speech.* This phonetic or articulatory plan, finally, can be executed by our articulatory system. *Articulation* involves the coordination of more than a hundred different muscles, creating the highly overlapping articulatory gestures that produce intelligible speech.

But there is more to speaking than producing overt speech. Speakers are always their own listeners. Just as we can listen to others and detect their intentions, hesitations, speech errors, we can parse our own speech and become aware of a less felicitous expression, a speech error or other problems of delivery. Here the mediating processing component is our own language comprehension system. In case of serious trouble the speaker may decide to stop and

make a self-repair. This is called *self-monitoring*. And that is another property of any complex skill.

How do these components cooperate in the generation of fluent speech? A first and important property of the system is what Kempen and Hoenkamp [4] have called *incremental production*. The high speed performance of the system (2-3 words per seconds, 10-15 phonemes per second in normal fluent speech) requires parallel processing, but it is of a special kind. Usually we don't plan the full message before we start speaking. The availability of a first partial notion to be expressed suffices to begin grammatical encoding; there is immediate activation of appropriate words and a beginning of sentence construction. And as soon as there is a sentence-initial word, phonological encoding is initiated. And we need only one or a few syllables and overt articulation can begin. Self-monitoring can begin as soon as any stretch of internal speech has become available. This roofing tile style of parallel processing is called incremental processing. It is simultaneousness in seriality. And it is only possible because of the *automaticity* of the various processing components. That, in fact, is a second major property of the system. We spend most of our attention on *what* to say, the topic of discourse, our communicative intention. But *how* we say it largely takes care of itself.

And finally, the system has a high degree of *modularity*. The various processing components do their own work largely independently of what the other components are doing at the same time. Our research shows that there is surprisingly little interaction or feedback in the system. And this makes good functional sense. The processing components involved have to perform wildly different tasks. If they were to affect each another during processing, the system would become highly error prone. But it isn't. Speech errors are rare events, say once in every several hundred words. Modularity is nature's protection against processing failure.

Let us now turn to the main theme of this paper, lexical access.

II. AN OVERVIEW OF WORD PRODUCTION

Figure 1 presents the stages in the production of words. Each stage belongs to a particular component of speech production mentioned above.

There is, first, conceptual preparation, yielding a lexical concept, usually as part of a larger message. Next there is lexical selection, retrieving the appropriate lexical item or lemma from the mental lexicon. This is an initial step in grammatical encoding; the retrieved word's syntax is a major determinant of further morpho-syntactic planning. Third, there is phonological encoding of that item in its prosodic context; this yields a so-called "phonological word", which may consist of more (and sometimes less) than a single lexical item. During the fourth stage, phonetic encoding, articulatory programs are created for each of a phonological word's syllables; these gestural programs have free parameters that are under the influence of procedures that compose phonological and intonational phrases. I will argue that the generation of syllabic

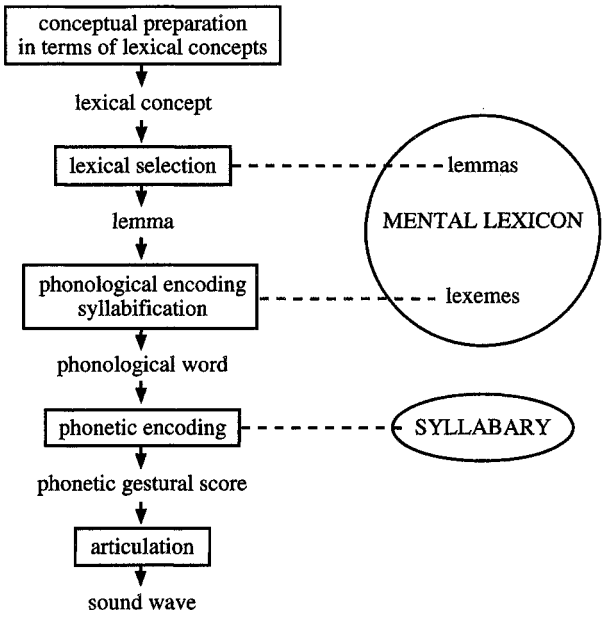

Fig. 1. Stages in the production of words

articulatory gestures involves access to a mental *syllabary*. In the final stage these gestures are executed by the respiratory, laryngeal and supralaryngeal structures of the vocal apparatus, such producing the overt articulation of a word. I will now discuss these stages in turn.

III. THE GENERATION OF LEXICAL CONCEPTS

A first step in word production is the activation of a lexical concept that captures your intention. This involves a process that I have called "perspective taking", which is part of micro-planning. It is easily exemplified by Figure 2.

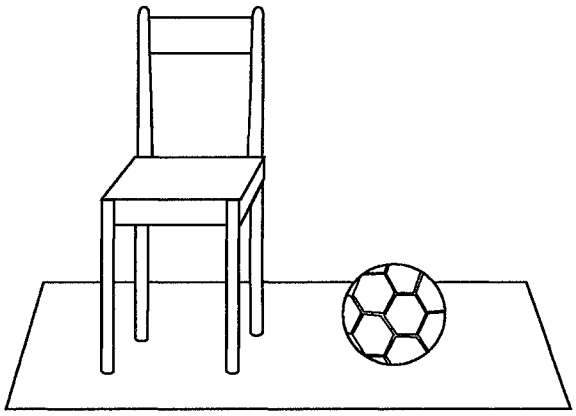

Fig. 2. Different perspectives on the same spatial relation.

When you watch the spatial scene in Figure 2, you may conceive it as a chair with a ball to the right of it. Though that would be a good message to formulate, it is by no means the only possible one. You may also conceive it as a ball with a chair to the left of it. Many of us can also

conceive it as a chair with a ball to the *left* of it. Some of us prefer to conceive it as a chair with a ball to the North or East of it, and there are more possibilities. Each of these involves a different perspective on the part of the speaker (for details, see [1][5]). There is no fixed relation between a referent (in this case a spatial relation) and the lexical concept that will capture it (LEFT, RIGHT, NORTH or otherwise). In fact, this relation is presently the subject of much scientific attention.

IV. LEXICAL SELECTION

Lexical selection is retrieving a word from the mental lexicon, given a lexical concept to be expressed. There is good reason to suppose that retrieving a word from the mental lexicon proceeds in two steps: selecting the semantically appropriate *lemma* and retrieving the corresponding word form or *lexeme* (a distinction originally proposed by Kempen en Huijbers [6]; see [7] for a review of the evidence). These two steps are depicted in Figure 1. Dell [8] was the first to propose an activation spreading model for this process. Here I will follow the activation spreading model that Roelofs of our Institute proposed [9].

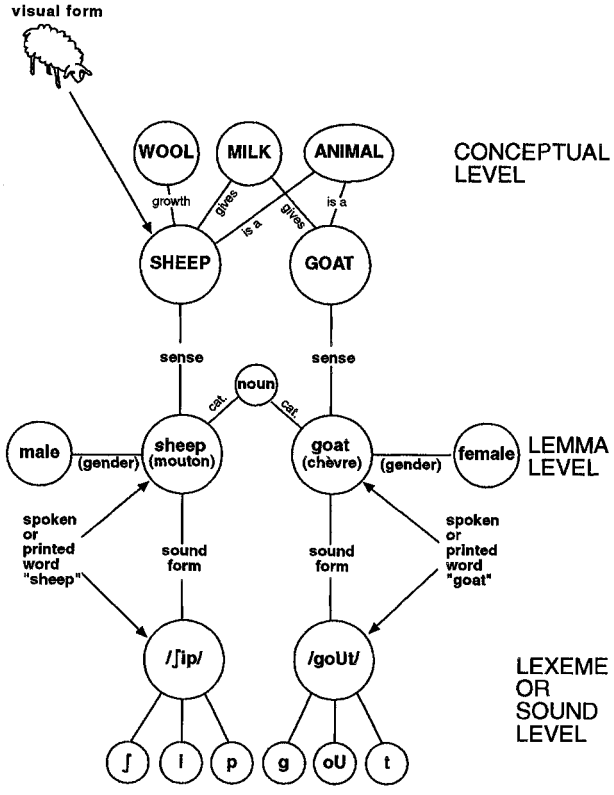

Fig. 3. Fragment of a lexical network. Note that the arrows represent types of connections, not the flow of information during production or comprehension. (reproduced from [10]).

A fragment of the network is presented in Figure 3. The network has three layers. The top, conceptual layer contains nodes that represent concepts, among them lexical concepts (such as the concept of SHEEP). The previous section was about this level of processing. One way of activating a lexical concept is to ask a subject to name a picture, for instance the picture of a sheep (see Figure 3). Strictly speaking, this layer is not part of the lexicon. But as soon as there is an active lexical concept node, activation spreads to the mental lexicon, initially to the lemma level. Nodes at this level, lemma nodes, represent a word's syntax: its category (noun, verb, adjective, etc.), a noun's gender (if any), a verb's subcategory structure, etc. Lexical selection is retrieving the appropriate lemma, to which I will presently return. After selection of a lemma, its activation spreads further to the third, lexeme level. Here nodes represent a word's phonological form. This information is used in the further step of phonological encoding (see below).

Roelofs proposed a simple rule for lexical selection. During any shortest time interval the probability of selecting a particular lemma is the so-called *Luce ratio*: the lemma's activation divided by the sum activation of all other lemmas. This probabilistic rule makes it possible to account for occasional semantic selection errors (such as *goat* for *sheep*), because activation spreads through the conceptual network (for instance SHEEP → ANIMAL → GOAT in Figure 3) and from any lexical concept down to its lemma. However, the model was tested by means of picture naming experiments. Here the subject was instructed to name a picture and to ignore acoustically or visually presented distractor words. These distractors could be semantically related to the picture (like distractor "goat" when the picture was one of a sheep) or unrelated. These manipulations affected naming latencies exactly as the model predicted.

V. PHONOLOGICAL ENCODING

There are four processes involved in the phonological encoding of words: *segmental spell-out, metrical spell-out, metrical word formation* and *segment-to-frame association.* This set of four processes, to be shortly discussed, begins by retrieving the lexeme (see Figure 3).

5.1 Word frequency

Jescheniak and Levelt [11] showed that the so-called *word frequency effect* arises at precisely this level, retrieving the lexeme. Pictured objects with high-frequent names (such as *boat*) are named faster than those with low-frequent names (such as *broom*). A main argument for locating the cause of the word frequency effect at the level of accessing lexemes, not lemmas or lexical concepts stems from experiments with homophones. Two different words are homophones when they are pronounced the same, for example *we* and *wee*. In the network model of Figure 3 homophones have different lemmas (*we* and *wee* even differ in syntactic category), but share the lexeme. Two homophones can have very different word frequencies (as is the case for *we* and *wee*, the former being very high-frequent, the latter very low-frequent) If the word frequency effect arises at accessing the lexeme, which is shared between homophones, the low-frequent partner should be as accessible as the high-frequent

partner. In fact, the low-frequent homophone (like *wee*), should behave as if it were high-frequent (i.e., like *we*). That counterintuitive result was indeed obtained in the experiments.

5.2 Segmental spell-out

The phonological information that comes available when a lexeme is accessed consists of (at least) two kinds, segmental and metrical information. The segmental information relates to the word's phonemic structure: its composition of consonants, consonant clusters, vowels, diphthongs, glides, etc. Theories differ with respect to the degree of specification, ranging from minimal or underspecification to full phonemic specification. Leaving this issue apart (but see [12]), a speaker who is in the process of producing the utterance *Police demand it*, will "spell out" the segments of the three words involved; let us limit to the latter two, which make up the syntactic verb phrase *demand it*. The spelled-out segments are /d/, /i/, /m/, /æ/, /n/, /d/ and /ɪ/, /t/. In several experiments from our laboratory it has been shown that segmental spell-out can be experimentally primed, leading to faster naming times (see for instance [13]).

5.3 Metrical spell-out

A lexeme's metrical information is the word's "frame", at least consisting of the word's number of syllables and its accent structure, the stress levels of successive syllables. When we are in the so-called "tip-of-the-tongue-state", we can often "sense" the metrical structure of the word that we cannot get hold of. Paul Meyer from our laboratory (unpublished doctoral dissertation) has shown that a word's metrical structure can also be primed. A picture of a cigar is named faster when subjects hear the distractor word *saloon* than when they hear the distractor word *salmon*. The former has the same (iambic) accent structure as the target word *cigar*.

The spelled-out metrical structures of our example words *demand* and *it* can be formally symbolized as σσ' and σ, respectively.

5.4 Metrical word formation

When our speaker utters *demand it*, the pronoun is cliticized to the verb; together they form one *phonological word*. Cliticization can only occur when there is no major syntactic boundary between the head word and the clitic. But interestingly enough, there are no *segmental* conditions on cliticization. The phonemic composition of the two words involved is never a restriction on their potential to cliticize. The process is an entirely metrical one, consisting of the fusion of two metrical frames, independent of the segmental content. In the case of our example a new metrical frame is composed out of the two spelled-out frames for *demand* and *it*. Formally:

$$\sigma\sigma' + \sigma \rightarrow \sigma\sigma'\sigma$$

Here the resulting tri-syllabic frame is the metrical structure of the phonological word *demandit*.

5.5 Segment-to-frame association

The final and crucial process in phonological encoding is the successive attachment of the spelled-out segments to the phonological word frame. In the example, the spelled-out segments /d/, ..., /t/ are successively associated to the metrical word frame σσ'σ. This process of association creates, "on the fly", the word's phonological syllables (in the example *de-man-dit)*, which may straddle lexical word boundaries (as does the last syllable, *dit*, in the example). In short, the phonological word, *not* the lexical word is the domain of syllabification in speech production. The process of on-line syllabification follows language-specific rules. For English a proposal can be found in [12].

It is possible to track the time-course of segment-to-frame association experimentally. Wheeldon and Levelt (forthcoming) had Dutch subjects listen to English words, with the instruction to internally monitor the Dutch translation equivalent of the English word for the occurrence of a prespecified target phoneme. For instance, the subject (who had good control of English as a second language) was given /f/ as a target phoneme to be monitored in the experiment. On presentation of the English word *hitch hiker*, the subject began internally generating the Dutch translation word *lifter*. As soon as /f/ would appear in the subject's generation of the phonological word, the subject would push a button, and we measured the reaction times. Phoneme monitoring reaction times (corrected for the latencies of comprehending the English word) are presented in Figure 4, as exemplified for the target word *lifter*. In the experiment subjects were given different target phonemes for the same words. For the target word *lifter* they were /l/, /f/, /t/, and /r/.

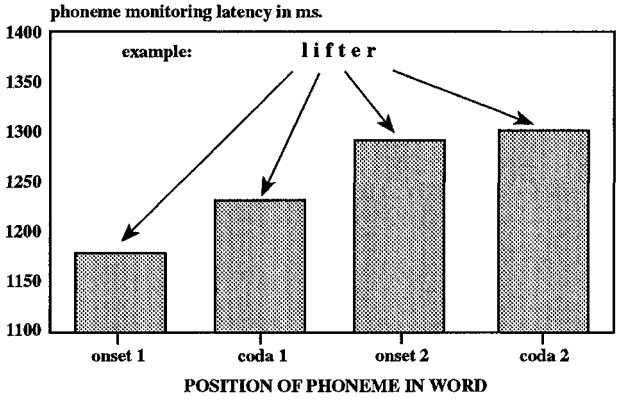

Fig. 4. Phoneme monitoring in phonological encoding. Reaction times for consonant targets in bisyllabic words: onset and coda of first syllable, onset and coda of second syllable.

Phoneme monitoring latencies increase significantly from left to right. The average latency difference between the onset of the first and of the second syllable is about 100 ms. This number should be compared to the time it takes to get from first syllable onset to second syllable onset in the overt articulation of words such as *lifter*. Those articulatory syllable durations are at least twice as long, i.e. over 200 ms. Apparently, phonological encoding is a

quite rapid process. Speech rate limitations are probably due to articulatory motor inertia rather than to processing limits in phonological encoding.

The eventual output of the stage of phonological encoding (see Figure 1) is a string of phonological syllables. In fluent speech they can appear as fast as one every 100 ms.

VI. PHONETIC ENCODING

The next stage of word formation in speech production is phonetic encoding (see Figure 1). As phonological words are formed, one by one, during phonological encoding, we retrieve the corresponding *phonetic* or *articulatory* syllables from our *mental syllabary*. At least, that is the present state of our theory on the production of words by Dutch speakers. In order to demonstrate the existence of a mental syllabary, Levelt and Wheeldon [12] argued as follows. We know that there is a word frequency effect, which arises in accessing a word's lexeme or phonological form (see above). If at a later stage we retrieve articulatory syllables (such as [*de*] when the phonological syllable */de/* is created, [*mæn*] when */mæn/* is created, etc.), these may also show a frequency effect. Infrequent syllables may come slower than frequent syllables. And if this is so, one can make the further prediction that the word and the syllable frequency effects are independent and additive, because they arise in successive, independent stages of word generation.

We tested this theory by having subjects produce bisyllabic words that were either high- or low-frequent as words, and that (orthogonally) varied in the frequency of their composing syllables. For instance, *lady* is a high-frequent word with high-frequent syllables, *language* is a high-frequent word containing low-frequent syllables, *litter* is a low-frequent word with high-frequent syllables and *lantern* is a low-frequent word consisting of low-frequent syllables. Figure 5 presents the naming latencies for these four categories of words.

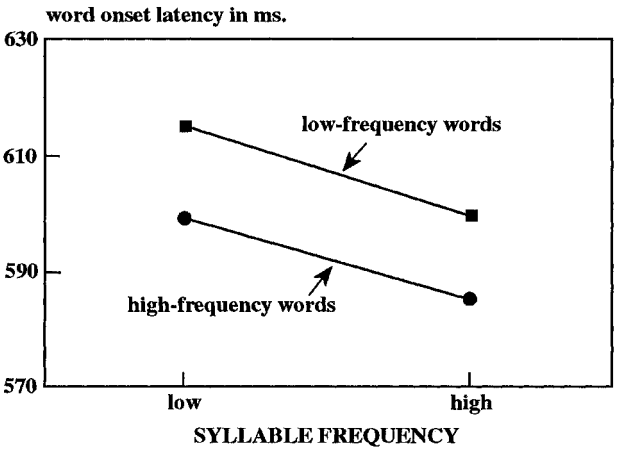

Fig. 5. Naming latencies for high- and low-frequency words that consist of high- or low-frequent syllables.

As expected, we found the usual word frequency effect (upper versus lower curve). But in addition we did find a significant syllable frequency effect (downward slope from left to right). And the two effects were independent and additive (the two curves are parallel). Further experiments showed that the syllable frequency effect is entirely due to the frequency of the word-final syllable (in accordance with the details of the processing theory proposed in [12]), and that it is not due to syllable complexity.

Although these data are in full agreement with the syllabary theory, it is still an open issue whether the findings can be replicated for other languages than Dutch. But there is hope. Dutch, like English has over 10.000 different syllables. If speakers can have a syllable store of this size, one should be optimistic about native speakers of languages such as Japanese or Chinese, who have to deal with no more than a few hundred syllables.

What is an articulatory syllable? It is a specification of the articulatory task that will produce the syllable, a *gestural score* in the sense of Browman and Goldstein [14]. This is a representation of the "tasks" to be performed at different articulatory tiers, the glottal, the velar and the oral tier. It is not a fully specified motor program (see [14] for details and further references).

The syllabary theory may give a new twist to the problem of underspecification. Above I mentioned that the segments that are spelled out during phonological encoding may be "underspecified" (as proposed by Stemberger [15]). For instance /k/ may not be specified for voicing when it appears in the context /s-r/ (as in the word *scruffy*). It can only surface as [k], not as [g]. But then, somewhere in the process, the segmental specification has to be completed. The syllabary theory may handle this problem in the following way. There is no need to complete the specifications of successive segments if one condition is met. It is that each phonological syllable arising during phonological encoding corresponds to one and only one gestural score in the syllabary. In that case, even if a syllable's segments are underspecified, their combination can still be unique.

The eventual output of phonetic encoding, then, is a string of gestural scores for successive syllables in a phonological words. Here we should add that these scores must have metrical and intonational parameters that proceed from higher-order prosodic processes, which are not discussed in this paper (but see [1] for details).

VII. ARTICULATION

The final stage in producing words consists of the execution of the gestural scores. Gestural scores underspecify the motor programs. If the score for the syllable [*pa*] specifies lip closure on the oral tier, that task may still be executed in a variety of ways, such as moving the upper lip, the lower lip, the jaw, or all three to some extent. The coordination structure theory developed by Saltzman and Kelso [16] (among others) accounts for the way in which the articulatory system performs the parameter reduction that is required to solve this mapping problem. The theory performs model-referenced control,

which involves an optimization of parameter choice, given the prevailing physical contingencies. The resulting least-effort solution takes into account such circumstances as the starting position of the articulators and physical contingencies such as having something in the mouth, the momentary state of bending of the vocal tract, etc. There are respectable alternative theories around as well (see [1] for a review), but since my Institute has not contributed to any of them, I will not go into details.

The final result of articulation is the overt production of words in the context of the utterance as a whole. And the speaker can monitor this output for occasional errors of delivery, and make a repair when necessary [17].

REFERENCES

[1] W. J. M. Levelt, "Speaking. From intention to articulation," Cambridge, MA: MIT Press, 1989.

[2] R. Jackendoff, "Consciousness and the computational mind," Cambridge, MA: MIT Press, 1987.

[3] S. N. Sridhar, "Cognition and sentence production: A cross-linguistic study," New York: Springer, 1988.

[4] G. Kempen & E. Hoenkamp, "Incremental sentence generation: Implications for the structure of a syntactic processor," J. Horecky (Ed.), Proc. 9th Int. Conf. Comp. Ling. Amsterdam: North-Holland, 1982.

[5] W. J. M. Levelt, "Perspective taking and ellipsis in spatial description," Manuscript MPI accepted for publication.

[6] G. Kempen & P. Huijbers, "The lexicalization process in sentence production and naming: Indirect election of words," Cognition, 14, pp. 185-209, 1983.

[7] W. J. M. Levelt et al., "The time course of lexical access in speech production," Psych. Review, 98, pp. 122-142, 1991.

[8] G. S. Dell, "A spreading activation theory of retrieval in sentence production," Psych. Review, 93, pp. 283-321, 1986.

[9] A. Roelofs, "A spreading activation theory of lemma retrieval in speaking," Cognition, 42, pp. 107-142, 1992.

[10] K. Bock & W. J. M. Levelt, "Language production: Grammatical encoding," M. A. Gernsbacher (Ed), Handbook of Psycholinguistics, San Diego: Acad. Press, 1994.

[11] J. Jescheniak & W. J. M. Levelt, "Word frequency effects in production: Retrieval of syntactic information and of phonological form. J. Exp. Psychol. LMC (in press), 1994.

[13] W. J. M. Levelt & L. Wheeldon, "Do speakers have access to a mental syllabary?," Cognition, 50, pp. 239-269, 1994.

[14] C. P. Browman & L. Goldstein, "Representation and reality: Physical systems and phonological structure,"Haskins Lab. Status. Rep. on Speech Res., SR-105/106, pp. 83-92, 1991.

[15] J. Stemberger, "Speech errors and theoretical phonology: A review," Bloomington: Indiana Ling. Club, 1983.

[16] E. Saltzman & J. A. S. Kelso "Skilled actions: A task-dynamic approach," Psych. Review, 94, pp. 84-106, 1987.

[17] W. J. M. Levelt, "Monitoring and self-repair in speech," Cognition, 14, pp. 41-104, 1983.