The ability to speak: from intentions to spoken words

Willem J. M. Levelt*

In recent decades, psychologists have become increasingly interested in our ability to speak. This paper sketches the present theoretical perspective on this most complex skill of *homo sapiens*. The generation of fluent speech is based on the interaction of various processing components. These mechanisms are highly specialized, dedicated to performing specific subroutines, such as retrieving appropriate words, generating morpho-syntactic structure, computing the phonological target shape of syllables, words, phrases and whole utterances, and creating and executing articulatory programmes. As in any complex skill, there is a self-monitoring mechanism that checks the output. These component processes are targets of increasingly sophisticated experimental research, of which this paper presents a few salient examples.

Talking is one of our favourite pastimes. Most of us spend large parts of the day in conversation, in teaching, in making phone calls, and so on. If we are not talking to others, we are likely to be talking to ourselves. The ability to speak is a gift of evolution to mankind. No other animal talks, whereas all healthy members of our species will eventually be able to talk. This skill is universal to our species and it must have played a key role in the survival of human society in the course of evolution. Language is the basic tool in co-operative action, in education, in the creation and transmission of culture and, quite generally, in the regulation of human bondage.

These obvious facts contrast sharply with the traditional lack of interest that psychology has shown in the subject of speaking. Take any of the classical textbooks in psychology or any of the myriad books on the psychology of everyday life, and the chances are that you will find no chapter or even section on speaking or conversation. The sparse exceptions, such as Wundt's textbook of 1896[1] or Freud's treatise on everyday psychopathology of 1904,[2] prove the rule. The analysis of how we speak was largely left to linguists, phoneticians and neurologists and they did an excellent job. When psychologists finally turned to their forgotten child some three decades ago, they found a wealth of linguistic theory, detailed analyses of speech errors, phonetic accounts of articulation and detailed aphasiological models of language disorders.

This formed an excellent basis for the development of a psychological perspective on this core human skill. More than in any of the sister disciplines, the psychological focus is on the *process* of speaking. How do you get from some communicative intention to the overt articulation of speech? What are the processing components involved in this generation of fluent speech and how do they interact in real time? However, to analyse this process, it is essential to know *what* it is that gets produced at different levels of processing. One has to understand semantic, syntactic, lexical and phonetic structure, as well as the breakdown of

* Max-Planck-Institut für Psycholinguistik, Wundtlaan 1, NL-6525 XD, Nijmegen, The Netherlands.

CCC 1062-7987/95/13-23

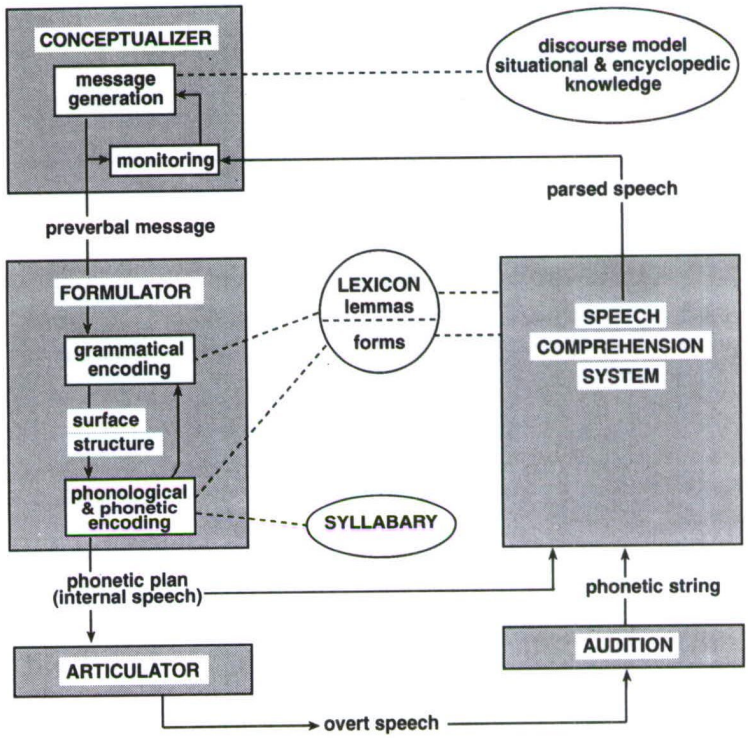


Figure 1. A diagram of the processing components involved in the generation of speech.

these structures after neurological damage, in order to design processing theories of any sophistication. When psychologists revived their interest in the process of speaking, the sister disciplines had already provided a firm knowledge base of this kind. Here, the new situation—some three decades ago—differed substantially from Wilhelm Wundt's and Karl Wernicke's over a century ago, when processing theories had to be developed almost from scratch.

Analysing the process of speaking requires accurate measurement of its time course. How fast do we retrieve an appropriate word, do we build a phrase around it, do we construct the articulatory gestures for successive syllables in the utterance, etc? To answer such questions it is necessary to measure a speaker's reaction times, for instance in naming objects. The generation of speech from thought is not instantaneous. Mental processes take time, as Donders argued long ago,[3] and we still use his methods of *mental chronometry* to trace and dissect mental events 'on the fly'. In recent years, mental chronometry has been successfully applied to the generation of speech, leading to a deeper understanding of this complex skill.

In the following I will present a global sketch of this skill's organization as we presently see it. In going over its constituent components, I will present occasional examples of experimental methods and findings. My aim here is to be informative (about some core issues and research methods) rather than comprehensive. For reviews of the field and of recent developments, see References 4, 5 and 6.

The diagram in Figure 1 is my summary view of what we presently know theoretically and empirically about the mental mechanism that generates speech from communicative intentions. At the same time, it is a working model for further research. According to this 'blueprint', the ability to speak is based on the interaction of a set of processing components that are relatively autonomous or 'modular' in their functioning. Each component is comparatively simple; the system's intelligence derives from the co-operation of the components. The blueprint is a working model in the sense that it raises three empirical issues. Is the partitioning of components correct? What operations are performed by each component? How do the components interact to generate fluent speech?

These issues are, of course, not independent. To model one component, one must keep an eye on the system as a whole.

Conceptual preparation

Our present state of knowledge differs for the different components. The situation is worst for the component labelled 'Conceptualizer'. Talking is an intentional activity, a way of acting. An effective utterance is one that makes the speaker's communicative intention recognizable to a partner-in-speech. And intentions are very diverse. The speaker may want to inform the interlocutor about something, to refer to something in the environment, to commit the interlocutor to some immediate or future action, to invite the interlocutor's sympathy, etc. Conceptualizing is primarily deciding on what to express, given the present intention. As speakers we spend most of our attention on these matters of content. There is much strategy here and rhetoric, involving politeness, wit, metaphor or, alternatively impudence, deceit, sarcasm and cynicism. In spite of much, and often highly creative, research (see for instance Reference 7), these mechanisms are not well-understood. It is certainly an oversimplification to lump them all together and suggest that the choice of content is a modular, coherent process that can be isolated from the rest of human intelligence.

However, there is more to conceptual preparation than the choice of content, and some of it has been successfully studied in recent years. One aspect is 'linearization' and another is 'perspective taking'. Both of these become relevant when the choice of content is being completed.

Linearization is the process by which we order information for expression. When somebody asks you to describe the layout of your home and you decide to comply, you have two problems to solve. Which features of your home will you mention—that is the content issue—and in what order? The latter is called the *linearization problem*. Your home is a three-dimensional structure, but speech is a linear medium. You can express just one thing at a time and the relevant items have to be ordered for expression. Speakers solve this problem by imagining a connected path through their living quarters as if they are taking you on a tour.

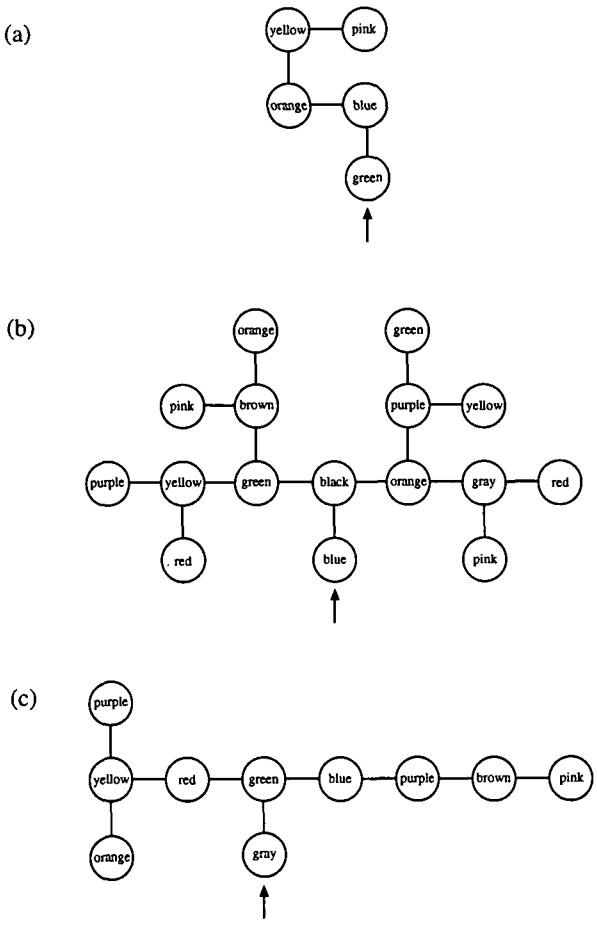


Figure 2. Patterns used to study how speakers order information for expression. Dots were coloured (here labelled by colour names).

How is such a path constructed? We studied this by asking subjects to describe patterns such as in Figure 2. The patterns consist of coloured dots, connected by arcs. Subjects were asked to begin their description at the dot marked by an arrow, and describe the pattern in such a way that the next subject would be able to draw the pattern from the tape-recorded description. How do speakers linearize such patterns? It turned out that linearization followed a few very simple rules. Speakers always follow a *connected* path through the pattern and register the choice nodes they pass and to which they have to return for completing their description. By the end of a path, they jump back to the most recent choice node on their 'memory stack' and begin a new path from there. This is recursively done until the whole pattern

has been covered. Which path do they choose when leaving a choice node? They strongly prefer to go for the simplest path first, in particular for the path that has the smallest number of choice nodes. This mechanism is so simple that the computed model consists of no more than a few lines. The mechanism, we could show, minimizes memory load (and hence attentional effort); it is fast and effective. It is also gullible. It is not hard to design patterns where nodes get skipped by the subject, for instance when they contain non-connected subgraphs. The same basic mechanism turns up in the other complex descriptions, for instance when you ask your subjects to describe their kin.

When you describe your apartment, or any spatial array for that matter, there is still more to be prepared. Every bit of spatial information that you want to express has to be given 'propositional shape', and this involves perspective taking. Assume your bathroom and kitchen are adjacent structures. You have a clear image of their arrangement in your place. You may now decide to express this imagistic information as *the bathroom is next to the kitchen*. Or do you prefer to say *the kitchen is next to the bathroom*? In some way or another you will locate one item with respect to the other item. There are many ways of doing this, but you will have to make some choice here. You will have to map geometrical and topological information onto logical (locative) relations between entities. There are preferences here that follow Gestalt principles. It is, for instance, more obvious to say *there is a chair in front of the cupboard* than to say *there is a cupboard behind the chair*. This is one aspect of perspective taking, the choice of referent and relatum. The preference is always to make the smaller, foregrounded object the referent that you locate with respect to the bigger, backgrounded object, the relatum.

But there is more to perspective taking. Figure 3 is another pattern that our subjects described. Two thirds of them described the directions with terms such as the ones on the outside of the figure: *up, up, right, right, down, left*. The pattern was flat on the table; yet still these subjects used terms such as *up* and *down*. What they are doing is making a gaze tour. They scan the pattern and tell you how their gaze moves: *up, up*, etc. We call this 'deictic' perspective, because the gaze indicates where you are in the pattern. If you turn the pattern by 90°

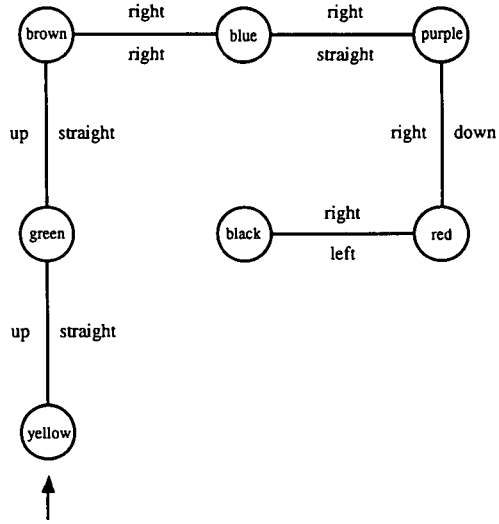


Figure 3. The use of direction terms depends on the speaker's perspective. On the outside, terms from a deictic perspective; on the inside, terms from intrinsic perspective.

or 180°, all direction terms will be different. One third of our subjects used the terms depicted on the inside of the pattern: *straight, straight, right, straight, right* and *right*. They make something like a body tour, as if they are walking or driving through the pattern, telling you what turns they take. These directions and turns only depend on the intrinsic shape of the pattern, not on the pattern's orientation with respect to the speaker; the terms will be the same when you turn the pattern by 90° or 180°. We call this 'intrinsic' perspective. But notice how different the terms are between the perspectives. The final move, for instance, is left in the deictic perspective and right in the intrinsic perspective. Does left mean right? Of course not. But the proposition that node X is left of node Y can simultaneously be true from one perspective and false from another perspective. What we see here is that the same spatial relation is captured by a different propositional relation, LEFT(X, Y) or RIGHT(X, Y), dependent on the chosen perspective. LEFT and RIGHT are lexical concepts, because we have words for them in the language, and to express some spatial relation in language we must translate it into lexical concepts. Perspective taking mediates here, and this is by no means limited to talking about spatial relations. To refer to any entity we must capture it by some lexical concept. I will refer to the same person as my

friend, colleague, brother, neighbour, or what have you, or to the same planet as morning star or evening star, dependent on the current chosen perspective in the conversation. Perspective taking is always at the core of a speaker's conceptual preparation.

The eventual result of conceptual preparation is some propositional structure that consists of lexical concepts. This is technically called the speaker's message. This conceptual structure is what the speaker will formulate, i.e. express in language.

Formulating

The Formulator performs two operations: grammatical encoding and phonological/phonetic encoding. Let us first attend to grammatical encoding.

Grammatical encoding

In order to encode a message linguistically, a first step must be to retrieve appropriate words for its lexical concepts (the word *left* for the concept LEFT, the word *right* for the concept RIGHT, etc.). As speakers we are equipped with a mental lexicon that contains tens of thousands of words. In normal conversation we retrieve words from this lexicon at a rate of 2 to 3 per second. That retrieval process is usually fully automatic; we do not have to spend much attention on it. I will shortly return to this process.

Each word we retrieve from the lexicon has a specification for the syntactic environment it requires. These environments are, for instance, different for verbs, nouns, adjectives, prepositions etcetera. But also within such classes there is great syntactic variety. During grammatical encoding the retrieved words are ordered and morphologically shaped in such a way that all these words' syntactic requirements are simultaneously met. This is a bit like solving a set of simultaneous equations, and there are sophisticated theories about how this process of syntactic unification proceeds.[8] Syntactic unification is fast and efficient, but also unintelligent in the sense that it ignores semantics. It only bothers about syntax. Hence, it can happen that we make speech errors such as this one:

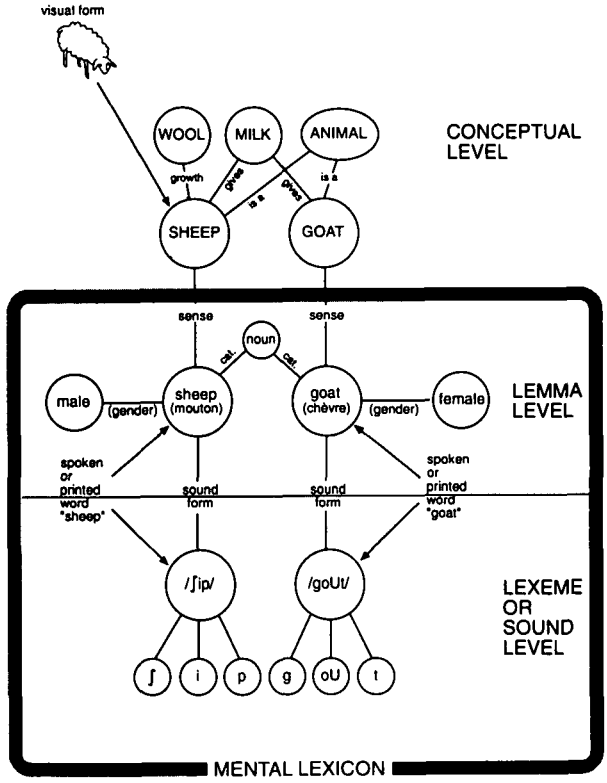


Figure 4. Fragment of lexical network.

Seymour sliced the knife with a salami

where the two nouns *knife* and *salami* exchanged positions, making the utterance very strange in terms of meaning. But the syntax is perfect. The grammatical encoder is a syntactic *idiot savant*.

How do we retrieve appropriate words to start with? In normal conversation we hardly ever make an error in selecting intended words. Selectional errors are below one in a thousand, on average. But they do occur, as in this example:

Don't burn your toes

Here the speaker intended to say *Don't burn your fingers*. Most selectional errors are semantic in nature, such as here where *toe* is selected instead of *finger*. How does this arise? Ardi Roelofs, of my laboratory, has developed a network model of lexical retrieval that can account for such semantic errors; Figure 4 shows a fragment of it. The nodes in the top layer represent lexical concepts. As discussed, these are the smallest or terminal

elements of a speaker's message. There is, for instance, a lexical concept for SHEEP and another one for GOAT. That these are semantically related is represented by their network of relations to other concepts. For instance, both are animals, both produce milk, etc. All this is encoded in the conceptual part of the network. When we ask a subject to name a picture and we show them one of a sheep, the lexical concept SHEEP will receive some activation.

How is the corresponding word retrieved from the lexicon? The theory says that an active concept spreads its activation to the lexicon. In the first instance, the activation spreads to the so-called lemma level of the mental lexicon. Nodes at this level, called lemmas, represent the syntactic properties of words. If SHEEP is the active concept, the lemma *sheep* and its syntactic properties will be activated. The lemma's activation spreads, for instance, to a node *noun*. In a language such as French that marks gender, the equivalent lemma *mouton* will spread activation to the gender node *male*, etc.

But if SHEEP is the active concept, some of its activation will spread (via ANIMAL, MILK, etc) to the semantically related concept GOAT. And if GOAT is active, it will activate its lemma *goat*. In the theory developed by Roelofs[9] the probability that at any given instant a particular lemma gets selected equals the lemma's activation divided by the sum activation of all lemmas at that moment. Since *sheep* receives much more activation that *goat* (via the big detour just described) the chances are that *sheep* gets selected, not *goat*. However, because the rule is probabilistic, there is always a minimal chance that there will be an error of selection. It is then likely to be a semantic error, because of the activation spreading in the meaning-based conceptual network.

Although the theory can explain this kind of error, it was tested and time and again confirmed in reaction time experiments. Here, we used the naming-interference method, introduced in our laboratory by Schriefers and Meyer.[10] The subject is shown a picture to be named as fast as possible. We measure the naming latency, i.e. the time from us presenting the picture to the subject initiating articulation. But we complicate this task by presenting a distractor word, acoustically or visually, that the subject is instructed to ignore. The distractor word can be presented

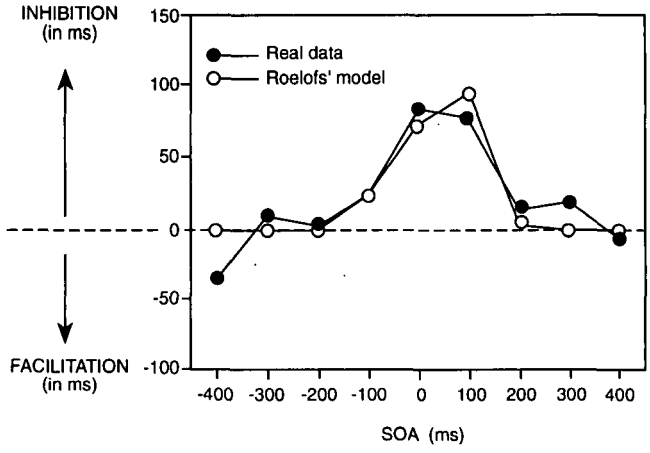


Figure 5. Reaction times from a picture–word interference experiment by Glaser and Düngelhoff and their fit by Roelofs's model.

simultaneously with the picture or a bit earlier or later, i.e. at different 'stimulus onset asynchronies', or SOAs. When the distractor word is semantically related to the target, for instance when we present *goat* as distractor when the picture is one of a sheep, the naming latency typically increases. It increases more than when we present an unrelated distractor (such as *house*). Figure 5 shows the classical results that Glaser and Düngelhoff[11] obtained with this paradigm. The naming latency is maximally affected when picture and distractor coincide, and it decreases for both longer and shorter SOAs. The figure also shows the (statistically perfect) fit of Roelofs's model to these data. Meanwhile, the model has been reconfirmed time and again in experiments from our own laboratory (see for instance References 9 and 12).

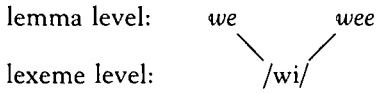
Phonological encoding

As soon as a lemma has been selected, its activation spreads to the next, lexeme level in the lexicon. At this level, a word's phonological properties are stored. For instance, when *sheep* is selected, its activation spreads to the corresponding lexeme /ʃip/, and from there to its individual segments /ʃ/, /i/ and /p/. These, in turn, are used to plan the articulatory shape of syllables, and I will shortly return to that process.

A preliminary point to consider, however, is

how we access the lexeme to start with. Here, an important phenomenon is the so-called word frequency effect. It is long known that we are relatively slow at naming objects that have names infrequent in everyday language use. For instance, it takes about 200 ms longer to name a picture of a moth than to name a picture of a mouth. In a picture recognition experiment, Joerg Jescheniak and I showed that this is not due to a difference in the ease of recognizing the objects. It is really due to accessing their names. The question then is this: does the word frequency effect arise in selecting the lemma or rather to accessing its lexeme, its word form information?

To study this we made use of homophones. Homophones are different words that are pronounced the same. An example is *we* and *wee*. In the network model of Figure 4 these words have different nodes at the lemma level; they differ in syntactic category (*we* is a pronoun, *wee* is an adjective), but they share a node at the lexeme level, because they are alike phonologically:



If the word frequency effect is caused at the lemma level, then it should differentially affect the two homophones of a pair. The high-frequency homophone, i.e. the pronoun *we* in the example, should be accessed faster than the low-frequency homophone *wee* in the example. That is the most likely event. However, if the effect is caused at the lexeme level, then one should observe something unexpected. The two homophones should be equally accessible, in spite of their frequency difference. More importantly, the low-frequency homophone (*wee* in the example) should behave as if it were a high-frequency word; it should be accessed just as fast as its high-frequency partner (*we* in the example). In our experiment[13] we had subjects produce low-frequency homophones (such as *wee*) as well as two types of control words: non-homophones that were just as low-frequency (for instance *vile*) and non-homophones that were just as high-frequency as the high-frequency homophone (for instance *me*, which is of about the same frequency as *we*). We wanted to know whether the low frequency homophone (like *wee*) behaves as if it is a low-frequency word (such as *vile*) or as if it is a high-frequency word (such as *me*).

The results of the experiment were unequivocal. Low-frequency homophones (like *wee*) are accessed just as fast as the high-frequency controls (like *me*), and they are a lot faster than the low-frequency control words (like *vile*). This proves that the frequency effect arises at the lexeme level. It is access to the word's phonological shape that is relatively slow for low-frequency words.

After accessing the lexeme, word form information becomes available, but not as a whole, as a complete word template. This has long been known from phonological speech errors, such as this one:

With this wing I thee red

Here the poor minister exchanged /w/ and /r/ in delivering this important formula. Psycholinguists have collected huge numbers of such errors, and the analyses show that a word's phonological segments are not fixated in their position, but have to be inserted in the right metrical slot as we speak. More specifically, a word's (or lexeme's) phonological information is of two kinds, the word's meter and the word's segments or phonemes. We know this intuitively from tip-of-the-tongue situations. When we can momentarily not recall a person's name, we often do know the name's meter or accent patterns (e.g. it is a three-syllable word with main stress on the first syllable) and can produce metrically similar names. Hence, the word's meter is independently retrieved. This also holds for normal, undisturbed retrieval. Paul Meyer of our institute showed this in a picture naming experiment. Subjects had to name pictures, such as one of a cigar. But when the picture was presented they also heard a distractor word, that they were instructed to ignore. The distractor word could have the same meter as the target or a different one. For *cigar*, for instance, distractors could be *saloon* (same iambic meter) or *salmon* (different trochaic meter). We had predicted that naming would be faster in the former case than in the latter, and that is what Meyer found.

This metrical information is an important ingredient in preparing connected speech. We do not speak in isolated words, but in larger metrical units. An important unit here is the phonological word. Let us consider an example. Somebody says *Police demand it*. Here *demand it* is a phonological word. How do we know? Word boundary information is lost in a phonological word, and that is

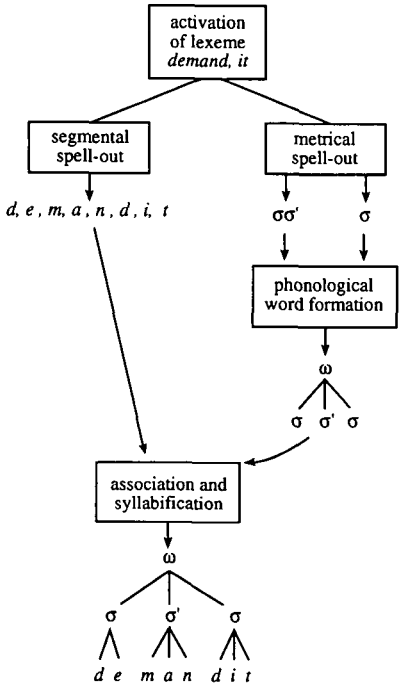


Figure 6. The encoding of a phonological word.

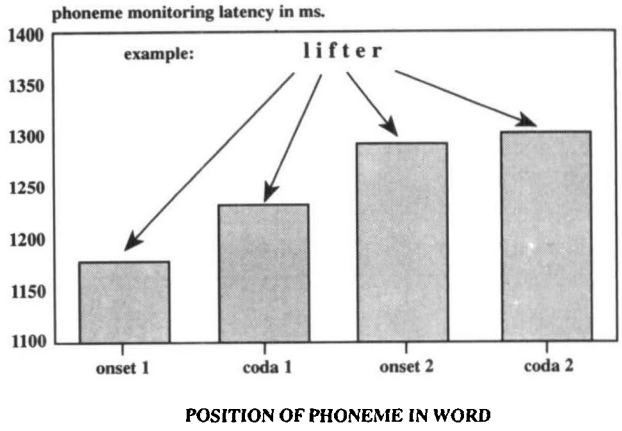


Figure 7. Detecting latencies for phonemes in internally generated translation words.

especially apparent in the word's syllabification. It is not the original lexical word (for instance *demand* or *it*) that is the domain of syllabification, but the phonological word. The eventual syllabification of our example will be *de-man-dit*; the last syllable *dit* straddles the lost boundary between the two composing lexical words.

We presently believe that phonological words such as *demandit* are generated as sketched in Figure 6. As soon as the metrical frames of *demand* and *it* have become available, they are blended to create a new three-syllable metrical frame. Then, the independently retrieved segments of *demand* and *it* are one-by-one associated to this new metrical frame, following rules that are more or less known.[14] As the association proceeds 'from left to right', syllables are created 'on the fly', first */de/*, then */man/* and then */dit/*. Antje Meyer and Herbert Schriefers obtained convincing experimental support for the theory that segments are associated with the metrical frame one after another, 'from left to right'.[15] These were picture–word interference experiments. Here I will mention another result that Linda Wheeldon and I recently obtained.[16] We gave our Dutch subjects a target phoneme to monitor, for instance the /f/ of *felix*. We then presented them with an English word whose Dutch translation they knew very well (most of our Dutch subjects have a good basic knowledge of English). For instance they heard the English word *hitch-hiker*. Its Dutch translation is *lifter*. The subject's task was to push a button if the Dutch translation contains the target phoneme. That is indeed the case for the translation equivalent of *hitch-hiker* (Dutch *lifter* has /f/ as its third segment). So, subjects never said *lifter*; they only pushed the button as soon as they became aware of the segment /f/ in the Dutch translation of *hitch-hiker*.

The results of this experiment, demonstrated from the word *lifter*, but based on a wide variety of target words, are presented in Figure 7. The figure shows the reaction time for /f/ as a target phoneme, but also for other target consonants in the word, /l/, /t/, and /r/. The reaction times increase significantly from left to right, supporting the notion that phonemes are indeed, one after another, attached to the metrical frame. Several control experiments excluded obvious alternative interpretations.

The main output, then, of phonological encoding is a string of phonological syllables, like */de/-/man/-/dit/*. But how do we generate the articulatory gestures for each of these syllables? This requires what we have called 'phonetic encoding'.

Phonetic encoding

The theory we are presently developing is that, as a speaker, you have access to a 'mental syllabary'

(see Figure 1). The syllabary contains a specification of the articulatory gesture for each phonological syllable you generate (i.e. one for /*de*/, one for /*man*/, one for /*dit*/, etc). The idea is that as soon as a syllable has been generated internally, its phonetic gesture is retrieved from the syllabary to be executed by the articulatory system.

How can one demonstrate the existence of such a syllabary in our minds? Levelt and Wheeldon[14] argued as follows. We know that there exists a word frequency effect and that it arises when the word form or lexeme information is retrieved (see above). If we then, at a later stage, retrieve syllabic patterns, another frequency effect may arise. We may be slower in retrieving a low-frequency syllable, one that is not used much, than a high-frequency one. Because these two retrieval steps (lexeme, syllable) are strictly sequential, the two frequency effects should be independent and additive. These are strong, non-trivial predictions. To test them we had subjects produce bisyllabic words of four kinds. The words could either be relatively high-frequency (such as *lady* and *language*) or relatively low-frequency (like *litter* and *lantern*). But each word either consisted of two high-frequency syllables (such as *lady* and *litter*) or of two low-frequency syllables (such as *lantern* and *language*). In other words, we completely crossed the two frequency variables. We measured the naming latencies.

The results are given in Figure 8 and they confirm our expectations precisely. Both frequency effects are statistically significant, but there is no interaction; the two effects are additive, the two curves are parallel. Various additional experiments showed that the syllable frequency effect is entirely due to the frequency of a word's last syllable. This is how it should be. The speaker initiates a word's pronunciation only after all of its syllables have been retrieved from the syllabary. However, each syllable retrieval is time-locked to completion in construing the corresponding phonological syllable, not to retrieving the previous syllable program. Hence, successive syllable frequency effects do not add; it is only the last syllable whose frequency effect we can observe in the latency data.

There is much more to phonological encoding than can be discussed here, such as the generation of intonation, of pauses, of loudness patterns, etc. Still, the preparation of syllables is at the core of speech preparation, because syllables are the basic units of articulation. The eventual output, then, of formulating is a phonetic or articulatory plan, which can phenomenologically present itself to us as internal speech.

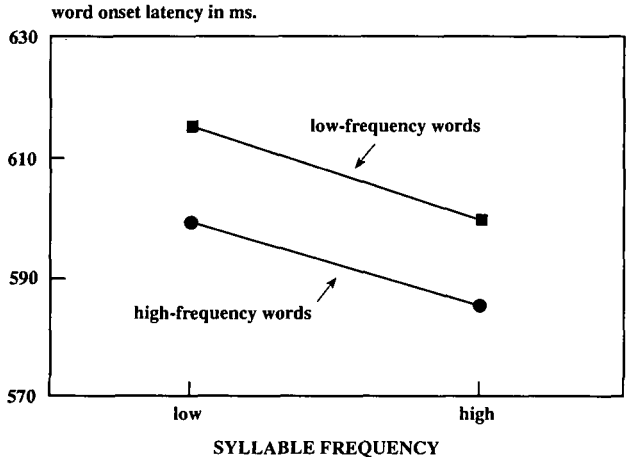


Figure 8. Naming latencies for words varying in word frequency and in syllable frequency.

Articulation

The phonetic plan, in particular the syllabic gestures, will eventually be executed by the articulatory apparatus, a highly sophisticated system that controls the movements of lungs, larynx, pharynx and mouth. A whole network of neural substrates in the cortex, the basic ganglia and the cerebellum are involved in this, our most complex motor behaviour. I will refrain from discussing this module (see Reference 4 for an extensive review), but still remark that this whole apparatus originally developed for us to breathe, to drink, to chew, to swallow. It is one of evolution's most impressive achievements that the same structures gradually developed to acquire a completely different function, the execution of speech. And this happened without abandoning the original functions. Surprisingly, this cohabitation turned out to be possible. The only concession we had to make is that we risk choking when we try to combine speaking and ingesting.

Self-monitoring

The person we listen to most is ourselves. We can listen to our internal speech, as we do when talking to ourselves. And we always listen to our own overt speech. That involves the same apparatus as listening to the speech of others. Just as we can detect errors, disfluencies or other problems of delivery in the speech of others, we can detect trouble in our own speech. If that trouble is serious enough, we can decide to halt and make a correction.

To study this process of self-monitoring, I collected and analysed some 1000 spontaneous self-corrections.[17] There are essentially three phases in the process of self-monitoring. The first phase I analysed was the halting process itself. What types of trouble make speakers halt? They are largely of two kinds. The first kind is all-out errors, as in *left to pink—er straight to pink*, where the subject made an error of lexical selection (*left* instead of *straight*). The second kind is an inappropriateness of sorts, where the speaker feels that further specification is necessary, as in *to the right is yellow, and to the right–further to the right is blue*. I also analysed how fast a speaker interrupts after the trouble appeared. The main rule here turned out to be quite simple. Halting is done right upon detecting the trouble, and this can be in the middle of a word. There is no tendency to safeguard the integrity of syntax in self-interruption, the break can be made anywhere in the sentence. But detecting can be slow, and the speaker will then stop one or more syllables after the trouble spot, as in *and from green left to pink–er from blue left to pink*, where *green* is the error.

The second phase is the editing phase. After halting speakers often use specific editing terms, such as *er-* to signal that trouble is on. It turned out that the editing term depends on the kind of trouble. Errors are mostly followed by terms such as *no, or* and *sorry*, whereas appropriateness trouble is predominantly signalled by terms such as *rather* or *that is*.

The third phase is the re-start, producing the self-correction proper. Where self-interruption fully ignores syntax, restarting is syntactically highly principled. Original utterance and repair relate in the same way as two conjuncts in a syntactic co-ordination. For example, *Is the nurse–the doctor interviewing patients?* is a normal well-formed repair, and so is the corresponding co-ordination *Is the nurse or the doctor interviewing patients?* But *Is the doctor seeing–the doctor interviewing patients?* sounds ill-formed, and so does the corresponding co-ordination *Is the doctor seeing or the doctor interviewing patients?* Notice that the repairs proper are the same in the two examples (*the doctor interviewing patients*). The crucial point is that the repair should syntactically fit the interrupted utterance. Apparently, in making a self-repair the speaker keeps the interrupted syntax in abeyance and grafts the correction onto it. This is, no doubt, the reason why it is often possible to 'splice away' self-corrections in recorded speech, a well-known practice in the broadcasting business.

Conclusions

The present paper briefly considered the 'blueprint of the speaker' in Figure 1. Its aim was to give a sketch of the main processing components that co-operate in the interaction of fluent speech and of the research methodology presently applied to the analysis of speech production. However, the blueprint was announced as a working model, as we do not have a comprehensive theory of this most complex of human skills, and it will be long before one is in the offing. What is new, however, is that we now have a theoretical perspective on this, our ability to speak. There is a satisfying (though of course not complete) agreement among colleagues in this field, on what the issues are and how they are mutually related. This is certainly an important condition for further progress. The other encouraging development is that we now have a wide range of empirical and, in particular, experimental methods by which the emergence of an utterance can be studied as a process. Mental chronometry has now firmly established itself in this field, together with a range of other methods that I did not have the opportunity to discuss. In short, this paper covered a subject of inquiry that will contribute essentially to our self-understanding as human beings, and that is now on the verge of becoming a coherent and successful enterprise.

REFERENCES

1. W. Wundt (1896) *Grundriss der Psychologie*. Kröner Verlag, Leipzig.

2. S. Freud (1904 (1954)) *Zur psychopathologie des Alltagslebens*. Fischer. Frankfurt am Main.

3. F. C. Donders (1869) 'Die Schnelligkeit psychischer Processe'. *Archiv Anatomie und Physiologie*, 657–681.

4. W. J. M. Levelt (1989) *Speaking. From Intention to Articulation*. MIT Press, Cambridge, MA.

5. W. J. M. Levelt (Ed) (1993) *Lexical Access in Speech Production*. Blackwell, Oxford.

6. Th. Herrmann and J. Grabowski (1994) *Sprechen. Psychologie der Sprachproduktion*. Spektrum. Heidelberg.

7. H. Clark (1992) *Arenas of Language Use*. University of Chicago Press. Chicago.

8. G. Kempen and J. Vosse (1989) Incremental syntactic tree formation in human sentence processing. *Cahiers de la Fondation Archives Jean Piaget*. Geneva.

9. A. Roelofs (1992) A spreading-activation theory of lemma retrieval in speaking. *Cognition* **42**, 107–142.

10. H. Schriefers, A. Meyer and W. J. M. Levelt (1990) Exploring the time course of lexical access in language production: picture-word interference studies. *J. Memory and Language*, **29**, 86–102.

11. W. A. Glaser and F. J. Düngelhoff (1984) The time course of picture-word interference. *J. Experimental Psychology: Human Perception and Performance*, **10**, 640–654.

12. A. Roelofs (1993) Testing a non-compositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, **47**, 59–87.

13. J. Jescheniak and W. J. M. Levelt (1994) Word frequency effects in production: Retrieval of syntactic information and of phonological form. *J. Experimental Psychology. Language, Memory and Cognition* **20**, 824–843.

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

15. A. Meyer and H. Schriefers (1991) Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *J. Experimental Psychology: Learning, Memory, and Cognition*, **17**, 1146–1160.

16. L. Wheeldon and W. J. M. Levelt (in press) Monitoring the time-course of phonological encoding. *J. Memory and Language*, **33**.

17. W. J. M. Levelt (1983) Monitoring and self-repair in speech. *Cognition*, **41**, 41–104.

Author's biography:

Willem J. M. Levelt is director of the Max Planck Institute for Psycholinguistics in Nijmegen and Professor of Psycholinguistics at Nijmegen University. He is a graduate of Leiden University and has held positions at the Institute for Perception in Soesterberg, Harvard University, the University of Illinois, Groningen University, the University of Louvain, and the Institute for Advanced Study in Princeton. Among his books are *On Binocular Rivalry* (1968), *Formal Grammars in Linguistics and Psycholinguistics* (1974) and *Speaking* (1989). He is a member of the Academia Europaea, the Royal Dutch Academy of Sciences and the Deutsche Akademie der Naturforscher Leopoldina.