Models of word production

Willem J.M. Levelt

How do we generate spoken words? This issue is a fascinating one. In normal fluent conversation we produce two to three words per second, which amounts to about four syllables and ten or twelve phonemes per second. These words are continuously selected from a huge repository, the mental lexicon, which contains at least 50–100 thousand words in a normal, literate adult person1. Even so, the high speed and complexity of word production does not seem to make it particularly error-prone. We err, on average, no more than once or twice in 1000 words2. This robustness no doubt has a biological basis; we are born talkers. But in addition, there is virtually no other skill we exercise as much as word production. In no more than 40 minutes of talking a day, we will have produced some 50 million word tokens by the time we reach adulthood.

The systematic study of word production began in the late 1960s, when psycholinguists started collecting and analyzing corpora of spontaneous speech errors (see Box 1). The first theoretical models were designed to account for the patterns of verbal slips observed in these corpora. In a parallel but initially independent development, psycholinguists adopted an already existing chronometric approach to word production (Box 1). Their first models were designed to account for the distribution of picture naming latencies obtained under various experimental conditions. Although these two approaches are happily merging in current theorizing, all existing models have a dominant kinship; their ancestry is either in speech error analysis or in chronometry. In spite of this dual perspective, there is a general agreement on the processes to be modeled. Producing words is a core part of producing utterances; explaining word production is part of explaining utterance production3,4. In producing an utterance, we go from some communicative intention to a decision about what information to express – the ‘message’. The message contains one or more concepts for which we have words in our lexicon, and these words have to be retrieved. They have syntactic properties, such as being a noun or a transitive verb, which we use in planning the sentence, that is in ‘grammatical encoding’. These syntactic properties taken together, we call the word’s ‘lemma’. Words also have morphological and phonological properties that we use in preparing their syllabification and prosody, that is in ‘phonological encoding’. Ultimately, we must prepare the articulatory gestures for each of these syllables, words and phrases in the utterance. The execution of these gestures is the only overt part of the entire process.

This review will first introduce the two kinds of word production model. It will then turn to the computational steps in producing a word: conceptual preparation, lexical selection, phonological encoding, phonetic encoding and articulation. This review does not cover models of word reading.

Two kinds of model

All current models of word production are network models of some kind. In addition, they are, with one exception5, all ‘localist’, non-distributed models. That means that their
The study of word production has two historical roots, one in speech error analysis and one in chronometric studies of naming.

The speech error tradition

In 1885, Meringer and Mayer published a substantial corpus of German speech error that they had diligently collected (Ref. 1). The corpus, along with the theoretical analyses they provided, is still a standard reference. An important distinction they made was between meaning-based substitutions (such as *ice* ‘(water)’ for *snow* ‘(snow)’) and form-based substitutions (such as *poper* (‘pet’) for *steife* (‘steamed’) or *mutter* (‘mother’) instead of *mutter* (‘mother’)). They showed that there is often a phonological connection in meaning-based errors (i.e. the over-representation of initial vowels). Freud was quick to discern the new generally accepted distinction between meaning- and form-based error by claiming that innate form errors are practically all meaning-driven (why does a patient say of her parents that they have *Giz* (‘good’) instead of *Geiz* (‘cheaper’)? Because she had suppressed her real opinion about her parents – oh, all the errors we would make!). A second, now classical distinction that Meringer and Mayer introduced was between exchanges (e.g. *weil* ‘(for) weil weude’) anticipated (‘toilet’ errors for *paple toilet*), perseverations (here again for *hore* and *hout* and *klee*), hesitations (often for *klee*), and omissions (omitted, blending, avoiding, and outside).

Many linguists and psychologists have continued this tradition (Ref. 1), but an hallmark consensus (probably triggered by the work of Cohen, Ref. 2) began in the late 1960s. In 1973, Finkel edited an influential volume of speech error studies, with part of her own collection of errors as an appendix (Ref. 3). Another substantial corpus was built up during the 1970s, the MIT-CU corpus. It led to two of the most influential models of speech production (Ref. 4). Garrett discovered that word exchanges, such as *like it and forget it behind* can span some distance and mostly preserve grammatical category as well as grammatical function within their clauses (Ref. 4). Sound-loan exchanges such as *took part for part* or *to part* and *to part* in *take part for part* on the other hand, ignore grammatical category and preferably happen between close words. This indicates the existence of two modular levels of processing in sentence production, a level where syntactic functions are assigned and a level where the ordering of forms (morphemes, phonemes) is organized. (2) Shattuck-Hufnagel’s scan-copier model concursive phonological sounding (Ref. 5). A core notion here is the existence of phonological frames, in particular syllable frames. Sound errors tend to preserve syllable position (as is the case in *back part*, or in *pape smoker for pipe smoker*). The model claims that a word’s phonemes are removed from the lexicon with their syllable position specified. They can only lend in the corresponding slot of a syllable frame.

In 1976, Baum, Merkey and MacKay (Ref. 6) developed a method for eliciting speech errors under experimentally controlled conditions, ten years after Brown and McNeil had created one for eliciting tip-of-the-tongue states (Ref. 7). Several more English-language corpora, in particular Stemberger’s (Ref. 8), were subsequently built up and analyzed, but sooner or later substantial collections of speech errors in other languages became available, such as Cohen and Nix’s homotax’s for Dutch (Ref. 8), Bogen’s (Ref. 9) for German, Garcia-Albea’s for Spanish (Ref. 10) and Ross and Porter-Ulanski’s for French (Ref. 11).

A final major theoretical tool in this research tradition was supplied by Dell (Ref. 12), who published the first computational model of word production, designed to account for the observed statistical distributions of speech error types.

The chronometric tradition

In 1895, Meringer and Mayer published a substantial corpus of German speech error that they had diligently collected (Ref. 1). The corpus, along with the theoretical analyses they provided, is still a standard reference. An important distinction they made was between meaning-based substitutions (such as *ice* ‘(water)’ for *snow* ‘(snow)’) and form-based substitutions (such as *poper* (‘pet’) for *steife* (‘steamed’) or *mutter* (‘mother’) instead of *mutter* (‘mother’)). They showed that there is often a phonological connection in meaning-based errors (i.e. the over-representation of initial vowels). Freud was quick to discern the new generally accepted distinction between meaning- and form-based error by claiming that innate form errors are practically all meaning-driven (why does a patient say of her parents that they have *Giz* (‘good’) instead of *Geiz* (‘cheaper’)? Because she had suppressed her real opinion about her parents – oh, all the errors we would make!). A second, now classical distinction that Meringer and Mayer introduced was between exchanges (e.g. *weil* ‘(for) weil weude’) anticipated (‘toilet’ errors for *paple toilet*), perseverations (here again for *hore* and *hout* and *klee*), hesitations (often for *klee*), and omissions (omitted, blending, avoiding, and outside).

Many linguists and psychologists have continued this tradition (Ref. 1), but an hallmark consensus (probably triggered by the work of Cohen, Ref. 2) began in the late 1960s. In 1973, Finkel edited an influential volume of speech error studies, with part of her own collection of errors as an appendix (Ref. 3). Another substantial corpus was built up during the 1970s, the MIT-CU corpus. It led to two of the most influential models of speech production (Ref. 4). Garrett discovered that word exchanges, such as *like it and forget it behind* can span some distance and mostly preserve grammatical category as well as grammatical function within their clauses (Ref. 4). Sound-loan exchanges such as *took part for part* or *to part* in *take part for part* on the other hand, ignore grammatical category and preferably happen between close words. This indicates the existence of two modular levels of processing in sentence production, a level where syntactic functions are assigned and a level where the ordering of forms (morphemes, phonemes) is organized. (2) Shattuck-Hufnagel’s scan-copier model concursive phonological sounding (Ref. 5). A core notion here is the existence of phonological frames, in particular syllable frames. Sound errors tend to preserve syllable position (as is the case in *back part*, or in *pape smoker for pipe smoker*). The model claims that a word’s phonemes are removed from the lexicon with their syllable position specified. They can only lend in the corresponding slot of a syllable frame.

In 1976, Baum, Merkey and MacKay (Ref. 6) developed a method for eliciting speech errors under experimentally controlled conditions, ten years after Brown and McNeil had created one for eliciting tip-of-the-tongue states (Ref. 7). Several more English-language corpora, in particular Stemberger’s (Ref. 8), were subsequently built up and analyzed, but sooner or later substantial collections of speech errors in other languages became available, such as Cohen and Nix’s homotax’s for Dutch (Ref. 8), Bogen’s (Ref. 9) for German, Garcia-Albea’s for Spanish (Ref. 10) and Ross and Porter-Ulanski’s for French (Ref. 11).

A final major theoretical tool in this research tradition was supplied by Dell (Ref. 12), who published the first computational model of word production, designed to account for the observed statistical distributions of speech error types.

The speech error tradition

In 1885, Meringer and Mayer published a substantial corpus of German speech error that they had diligently collected (Ref. 1). The corpus, along with the theoretical analyses they provided, is still a standard reference. An important distinction they made was between meaning-based substitutions (such as *ice* ‘(water)’ for *snow* ‘(snow)’) and form-based substitutions (such as *poper* (‘pet’) for *steife* (‘steamed’) or *mutter* (‘mother’) instead of *mutter* (‘mother’)). They showed that there is often a phonological connection in meaning-based errors (i.e. the over-representation of initial vowels). Freud was quick to discern the new generally accepted distinction between meaning- and form-based error by claiming that innate form errors are practically all meaning-driven (why does a patient say of her parents that they have *Giz* (‘good’) instead of *Geiz* (‘cheaper’)? Because she had suppressed her real opinion about her parents – oh, all the errors we would make!). A second, now classical distinction that Meringer and Mayer introduced was between exchanges (e.g. *weil* ‘(for) weil weude’) anticipated (‘toilet’ errors for *paple toilet*), perseverations (here again for *hore* and *hout* and *klee*), hesitations (often for *klee*), and omissions (omitted, blending, avoiding, and outside).

Many linguists and psychologists have continued this tradition (Ref. 1), but an hallmark consensus (probably triggered by the work of Cohen, Ref. 2) began in the late 1960s. In 1973, Finkel edited an influential volume of speech error studies, with part of her own collection of errors as an appendix (Ref. 3). Another substantial corpus was built up during the 1970s, the MIT-CU corpus. It led to two of the most influential models of speech production (Ref. 4). Garrett discovered that word exchanges, such as *like it and forget it behind* can span some distance and mostly preserve grammatical category as well as grammatical function within their clauses (Ref. 4). Sound-loan exchanges such as *took part for part* or *to part* in *take part for part* on the other hand, ignore grammatical category and preferably happen between close words. This indicates the existence of two modular levels of processing in sentence production, a level where syntactic functions are assigned and a level where the ordering of forms (morphemes, phonemes) is organized. (2) Shattuck-Hufnagel’s scan-copier model concursive phonological sounding (Ref. 5). A core notion here is the existence of phonological frames, in particular syllable frames. Sound errors tend to preserve syllable position (as is the case in *back part*, or in *pape smoker for pipe smoker*). The model claims that a word’s phonemes are removed from the lexicon with their syllable position specified. They can only lend in the corresponding slot of a syllable frame.

In 1976, Baum, Merkey and MacKay (Ref. 6) developed a method for eliciting speech errors under experimentally controlled conditions, ten years after Brown and McNeil had created one for eliciting tip-of-the-tongue states (Ref. 7). Several more English-language corpora, in particular Stemberger’s (Ref. 8), were subsequently built up and analyzed, but sooner or later substantial collections of speech errors in other languages became available, such as Cohen and Nix’s homotax’s for Dutch (Ref. 8), Bogen’s (Ref. 9) for German, Garcia-Albea’s for Spanish (Ref. 10) and Ross and Porter-Ulanski’s for French (Ref. 11).

A final major theoretical tool in this research tradition was supplied by Dell (Ref. 12), who published the first computational model of word production, designed to account for the observed statistical distributions of speech error types.
to the subject at different SOAs with respect to picture onset. The distracter words were either semantically or phonologically related to the target word, or unrelated. This paradigm and its many later variants made it possible to study the relative time course of the target name’s semantic and phonological encoding in much detail.

References

- Frankin V.A. (1972) Speech error as linguistic evidence, Minnetonka.
- Friesen, F. (1967) Latency of different verbal responses to the same stimulus J. Exp. Psychol. 70, 302–305
- Peter, M.G. et al. (1986) lexical and conceptual representation in beginning and proficient bilingual J. Verb. Learn. Verb. Behav. 22, 23–38
- Stroup, J. (1921) Studies of interference in serial verbal interactions J. Exp. Psychol. 9, 483–484

...
The main strata in this network are the same as those in the interactive model. There is a conceptual/semantic level of nodes, a lemma stratum and a phonological or form stratum. But the model is only partially interactive. There are good reasons for assuming that conceptual and lemma strata are shared between production and perception, hence their interconnections are modelled as bi-directional. But the form stratum is unique to word production; it does not feed back to the lemma stratum. Therefore it is often called the discrete (as opposed to ‘interactive’) two-step model. Although the model was designed to account for response latencies, not for speech errors, the issue of ‘mixed’ speech errors cannot be ignored and it has not been. The explanation is largely post-lexical. We can strategically monitor our internal phonological output and intercept potential errors. A phonological error that happens to create a word of the right semantic domain (such as rat for cat) will have a better chance of ‘slipping through’ the monitor than one that is semantically totally out of place (such as mat for cat). Similarly, an error that produces a real word will get through easier than one that produces a non-word. There is experimental evidence that the monitor is indeed under strategic control. Still, the causation of mixed errors continues to be a controversial issue among models of word production.

Conceptual preparation

The first step in accessing content words such as cat or select is the activation of a lexical concept, a concept for which you have a word or morpheme in your lexicon. Usually, such a concept is part of a larger message, but even in the simple case of naming a single object it is not trivial which lexical concept you should activate to refer to that object. It will depend on the discourse context whether it will be more effective for you to refer to a cat as cat, animal, siamese or anything else. Rosch has shown that we prefer ‘basic level’ terms to refer to objects (cat rather than animal; dog rather than collie, etc.), but the choice is ultimately dependent on the perspective you decide to take on the referent for your interlocutor. Will it be more effective for me to refer to my sister as my sister or as that lady or as the physicist? It will all depend on shared knowledge and discourse context. This freedom of perspective-taking appears quite early in life and is ubiquitous in conversation.
Working models of word production begin where perspective-taking ends: at the activation of a target concept to be expressed. The representation of a target concept, however, varies among models. The two preferred variants are just the ones exemplified in Figs 1 and 2. Concepts are either represented as decomposed, or as non-decomposed or ‘whole’. The issue is controversial, but arguments have been accumulating for using whole-concept representations in models of word production. One argument is the so-called ‘hyperonym problem’. If you activate some set of semantic features as a representation of the notion ‘cat’, the notion ‘animal’ will involve a proper subset of these features. Hence, it is indeterminate which of the two will ultimately be expressed. This is not an advantage: hyperonym speech errors are rare in any case and you need extra machinery to prevent the hyperonym problem from arising.
Lexical selection

In the contextual tradition lexical selection has been studied with interference paradigms, in particular picture-word interference (see Box 1). The recurring finding has been that naming an object is slowed down when a distract word is presented with the picture; the effect is stronger when the distract word is semantically related to the target than when it is semantically unrelated and it is at maximum when picture and distract word are presented simultaneously. The WEAVER model provides an accurate quantitative account of a wide range of picture-word interference data, with only a few free parameters. How does it work? When you are naming a picture of a sheep and you decide to go for the basic level term, you will activate the lexical concept 'sheep' as your target and activation spreads to the corresponding lemmas. In the semantic network activation spreads to related concepts, such as 'goat' and 'llama'. Then, in turn, spreading activation to their lemmas. During any unit time interval the probability of selecting the target lemma 'sheep' from the mental lexicon is the ratio of that lemma's degree of activation and the total activation of all lemmas (including 'goat', 'llama' and 'sheep'). This is called Luís's ratio, and it allows for the computation of an expected selection latency. In other words, there is competition between semantically related lemmas. Active alternatives slow down the selection process (even though a special checking mechanism in WEAVER normally prevents them from replacing the target). If you present the semantically related word 'goat' as a distractor, the already co-activated lemma 'goat' will receive an additional boost, thereby becoming a strong competitor to 'sheep'. By contrast, if you present a semantically unrelated word, such as 'chair', as distractor, there will be no convergence of activation and, correspondingly, competition will be relatively weak. That explains the semantic-inhibition effect.

Activation spreading through a semantic network (of whatever type) is also the obvious explanation for semantic naming errors, the dominant speech error type (about two-thirds of errors in a normal picture naming task are semantic in character). But what is a semantic error? A particular choice of words may have its cause in perspective-taking. They successfully modeled the naming errors (semantic and other) of a diverse set of aphasic patterns by manipulating no more than two parameters in their interactive two-step model: the weight on the network connections and the decay rate of the nodes' activation. When you produce a sentence, the moment of selecting the most activated lemma is dictated by when it is to be inserted in the grammatical frame. The selection moment is usually given a constant default value in modeling error distributions.

Both whole-concept and featural representations allow for precise semantic inferences (of the type ‘a dog is an animal’), but this inferential potential plays no role in the factual word production process.

---

**Fig. 3. Three steps in the morpho-phonological encoding of the word selecting.**

**Step 1. Accessing the morpho-phonological code**

The target lemma is select, marked for progressive tense. Two codes are successively accessed: first the code for the head morpheme select, then the code for the suffix morpheme -ing. For each code the speed of access is dependent on its frequency of usage.

**Step 2. Spelling out the phonological code**

Spelling out segments is one by one attached to the metrical code. Following the rules of the language, the onset of the first syllable and its nucleus or vowel are specified; that completes the first syllable. Then becomes the onset of the second syllable and its nucleus.

**Step 3. Prosodification**

The timing of lexical selection is not explicitly modeled in the chronometric tradition lexical selection has been studied with interference paradigms, in particular picture-word interference (see Box 1). The recurring finding has been that naming an object is slowed down when a distract word is presented with the picture; the effect is stronger when the distract word is semantically related to the target than when it is semantically unrelated and it is at maximum when picture and distract word are presented simultaneously. The WEAVER model provides an accurate quantitative account of a wide range of picture-word interference data, with only a few free parameters. How does it work? When you are naming a picture of a sheep and you decide to go for the basic level term, you will activate the lexical concept 'sheep' as your target and activation spreads to the corresponding lemmas. In the semantic network activation spreads to related concepts, such as 'goat' and 'llama'. Then, in turn, spreading activation to their lemmas. During any unit time interval the probability of selecting the target lemma 'sheep' from the mental lexicon is the ratio of that lemma's degree of activation and the total activation of all lemmas (including 'goat', 'llama' and 'sheep'). This is called Luís's ratio, and it allows for the computation of an expected selection latency. In other words, there is competition between semantically related lemmas. Active alternatives slow down the selection process (even though a special checking mechanism in WEAVER normally prevents them from replacing the target). If you present the semantically related word 'goat' as a distractor, the already co-activated lemma 'goat' will receive an additional boost, thereby becoming a strong competitor to 'sheep'. By contrast, if you present a semantically unrelated word, such as 'chair', as distractor, there will be no convergence of activation and, correspondingly, competition will be relatively weak. That explains the semantic-inhibition effect.

Activation spreading through a semantic network (of whatever type) is also the obvious explanation for semantic naming errors, the dominant speech error type (about two-thirds of errors in a normal picture naming task are semantic in character). But what is a semantic error? A particular choice of words may have its cause in perspective-taking. They successfully modeled the naming errors (semantic and other) of a diverse set of aphasic patterns by manipulating no more than two parameters in their interactive two-step model: the weight on the network connections and the decay rate of the nodes' activation. When you produce a sentence, the moment of selecting the most activated lemma is dictated by when it is to be inserted in the grammatical frame. The selection moment is usually given a constant default value in modeling error distributions.
Box 2. Implicit priming

The method of implicit priming was introduced by Meyer to study the time course of phonological encoding, that is the speaker's construction of a spoken word's form (Ref. a). The initial and major discovery, which has been repeatedly confirmed, was that a word's form is built up incrementally, starting with the first segment. Apparently, phonological word shapes do not come as whole templates; rather they are generated afresh, time and again, from beginning to end.

The method is exemplified in Table 1. Subjects learn a set of three semantic–word associations (A–B), for instance set 1 in the leftmost column. Then, an A-word from the set appears on the screen and the subject produces the corresponding B-word as fast as possible. The word onset latency is measured by voice key. The A-words from the set are repeatedly presented in random order and at each trial the naming latency of the B-word is registered. Then the subject is presented with set 2, the triple in the second column of the table below, and the same procedure is run for that set. Finally, set 3 is run in the same way.

The response words in a set share a phonological property. The B-words in set 1 are lower, bowl and tone; they share the initial syllable /l/. Similarly, the B-words in set 2 share the initial syllable /b/; and those in set 3 share the initial syllable /f/. Such sets sharing a phonological property are called "homogeneous" and the shared property is called the "implicit prime".

Can the subject use this implicit prime when running through the set? Whether the subject can prepare for the first syllable of the response word can be tested by comparing the homogeneous condition with a heterogeneous condition; that is, one in which there is no implicit prime. The heterogeneous condition is created by reordering the A–B pairs in such a way that they no longer share their first syllable. For instance, the first set of the heterogeneous condition (fourth column in the table) has lower, fence and major as response words. Each word pair is in its own column in the experiment: it appears in both the homogeneous and the heterogeneous condition.

In the heterogeneous condition there is no implicit prime, hence the subject cannot prepare anything. When Meyer did the experiment exemplified in Table 1 (in Dutch), she found that response latencies were significantly shorter in the homogeneous condition than in the heterogeneous condition. Apparently, subjects can prepare for the response word's first syllable if the first syllable of the response word can be tested by comparing the homogeneous condition with a heterogeneous condition; that is, one in which there is no implicit prime. The heterogeneous condition is created by reordering the A–B pairs in such a way that they no longer share their first syllable. For instance, the first set of the heterogeneous condition (fourth column in the table) has lower, fence and major as response words. Each word pair is in its own column in the experiment: it appears in both the homogeneous and the heterogeneous condition.

Homogeneous condition Heterogeneous condition

<table>
<thead>
<tr>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>place–local</td>
<td>signal–beacon</td>
<td>captain–major</td>
<td>single–toner</td>
<td>place–local</td>
<td>fruit–lotus</td>
</tr>
<tr>
<td>glass–beaker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


Morpho-phonological encoding

When you are planning the sentence 'they are selecting me', you must retrieve from your lexicon the morpho-phonological codes for each of the selected words, among them the two morpheme-size codes select and ing (see Fig. 3), and compute their syllabification and accent structure in context (so–de–ting). This naturally divides the process into 'code retrieval' and 'prosodification'.

Code retrieval

An item's morpho-phonological code consists of its morphological make-up, its metrical shape and its segmental make-up (see Fig. 3, Step 3 and Step 2). Retrieving that information follows activation/seletion of the lemma 9. Much ink and many subjects have been spilled over this issue. In the WEAVER model, the activation and retrieval of a phonological code is strictly conditional on selecting the corresponding lemma. For instance, when your target word is eat, you first select its lemma and only then spread activation to its phonological code (lat). This predicts that alternative active, but non-selected lemmas (such as the lemma for sit) do not spread any activation to their phonological codes.

Initial experimental evidence 9 showed that, in picture naming, there is semantic but indeed no phonological activation of same-category alternatives (if eat is the target, die is semantically but not phonologically active). All
speech-error based models of word production, however, assume that there is free cascading of activation throughout the network. Hence, active alternatives should also become active phonologically, at least to some extent24. However, the original finding was reconfirmed in a quite critical replication25. Still, evidence for phonological co-activation of semantic alternatives was obtained for one restricted case: if the alternative is a synonym of the target26,27. When you name the picture of a couch, the phonological code of _sofa_ is measurably co-activated. The cause of this robust finding is unclear. It shows that cascading exists, but not that it is a general property of the lexical network – after all, it doesn’t show up for same-category items such as cat and dog. I suggest that the phenomenon is related to perspective taking.

When you have two equivalent ways of making reference to an object, you may occasionally select both lemmas and hence spread activation to both phonological codes. This means that WEAVER’s special checking mechanism (see above) can occasionally fail if two highly competitive lemmata are about equally activated. There is suggestive speech error evidence that this indeed occurs: phonological word blends tend to be blends of near-synonyms (such as _close_ and _near_ blending into _clert_), hardly ever of same-category items.

There is a strong and robust word-frequency effect in word production (which is in part an age-of-acquisition effect)28. Controlling for conceptual biases, you are typically faster in producing a high-frequency response such as _cat_ than a low-frequency response such as _work_. It is now known that the effect arises in accessing the phonological code (Fig. 3, Step 3), not in selecting the lemma29. This fact has a suggestive relation to the so-called tip-of-the-tongue (TOT) phenomenon. It happens occasionally that, while normally speaking, you get stuck on the name of a person, flower, instrument or whatever. The phenomenon can be experimentally induced by presenting a subject with the definition of an object to be named. If the name is low-frequency, you often induce the TOT state in the subject. When the target language is gender-marking (such as Italian), there is a good chance that the subject knows the gender of the problem word30 and this also holds for the much amplified case of word finding trouble in many agrammatic patients31. This has been used as one of many arguments for the distinction between an ‘earlier’ syntactic lemma-level and a ‘later’ phonological code level in the lexical network32,33. But that argument has provoked some controversy34–36 which, so far, are unresolved. Probably more relevant speech error evidence for the precedence-of-syntax claim is the repeated finding of almost absolute gender preservation in phonological word substitution errors (such as _lasagna_ for _spaghetti_ in Italian)37–39. Most of these errors are real, on-line productions of the lexical network. So far, however, they have not been modeled.

Prosodification

The core process here is incremental syllabification. Let us return to the target sentence _they are selecting me_ (Fig. 3, Step 3). The morpho-phonological code of the progressive lemma _select_ consists of two morpho-phonological packages, _s, i, l, k_ and _t_. Syllabification proceeds ‘from left to right’. You first chunk the first two phonemes to create the syllable _si_. You then take the next three to compose the syllable _kt-_. Finally, you chunk the remaining segments to compose _i_ (the best evidence for the strict incrementality of this process comes from experiments using the ‘implicit priming paradigm’ (see Box 2). Notice that the last syllable, _l/kt/_, straddles two morphemes, _select_ and _ing_. This can also happen across words. When you utter _they will select us_ the syllabification will be _s-it-kt-i-ts-i/_ where _t/kt/_ straddles the words _select_ and _us_. But when you produce _they select me_, the syllabification is _s-it-kt-i-ts-i_, without straddling. Apparently, the syllables are not given in the phonological code of the morpheme, but deputed on the context in which the word and its morphemes appear. The word’s phonemes are not marked for a fixed position in their syllables; the _s_ in _select_ will appear as syllabic onset or syllable offset, dependent on the context. The domain of syllabification (such as selecting, selecting, _select_ is called the ‘phonological word’.) It can be larger or smaller than the lexical word. The incremental ‘chunking’ of segments in the on-line composition of syllables follows a strict set of rules, which vary among languages2. These rules are rapidly applied, then and again, in the fluent generation of speech. When you are a speaker of Papuan Hua, all your syllables consist of a consonant (C) followed by a vowel (V), CV. Other languages have one or more other syllable frames in addition, such as _V, CV,C, CCV_ and so on. Traditionally, syllabification was conceived of as filling such syllabic frames (see Box 1), but arguments for this view have become less convincing30–32. In particular, the idea that phonemes in the phonological code (such as _t/kt_ in _select_ are marked for a particular syllable position creates more problems than it solves. The preference of sound exchanges (such as _maggy_ become _maggy marly_) to preserve syllabic position can be explained differently, as a combination of word onset vulnerability, phoneme similarity and phonotactic restrictions.

There is good chronometric evidence, however, for the existence of morphal frames (see Fig. 3). For Dutch, and probably for other stress-assigning languages such as English and German, there is a dominant metrical pattern:

**Outstanding questions**

- How should error-based and chronometric models be further reconciled computationally and empirically?
- What causes a speech error? Is it caused by occasional cascading or occasional feedback in a normally non-cascading, feed-forward system?
- Is it the product of noise in a normally cascading interactive system? Or is the origin of speech error something else entirely?
- How does the word-production network relate to the word-perception network? Is it self-monitoring realized in this combined system?
- How are syllabic and larger gestures computed from a syllabified phonological code? Is there anything like a repository of syllabic gestural scores?
- If phonological word encoding is an incremental process, why is it that naming a short word is harder than naming a long word?
- Which brain regions subserve the core components of conceptual-semantic preparation, lexical selection, phonological code retrieval, prosodification, phonetic encoding, articulation and self-monitoring?
word stress goes to the first full-voweled syllable (morning, yellow, forget) – the ‘st’ in the latter word is not full-voweled, but rather a neutral ‘schwa’-sound. This can be automatically produced in incremental syllabification. But when a word has a deviant stress pattern, the automaticity breaks down14–15 (see Box 2 for an example). A word’s deviant metrical frame is probably stored as part of its phonological code; it guides the deviant prosodification. Languages differ, however, in their default metrics.

The distinction between accessing a word’s phonological code and in subsequent rapid syllabification is crucial for understanding the neural architecture of word production. A meta-analysis of imaging studies in word production16 suggests that accessing the code involves Wernicke’s area, whereas prosodification involves the posterior inferior frontal cortex.

Phonetic encoding and articulation

As incremental prosodification proceeds, the resulting syllabic and larger prosodic structures should acquire phonetic shape. As a speaker you will incrementally prepare articulatory gestures for the syllables in their prosodic context. A core feature of the WEAVER model is the notion of a syllabic score17. Statistics show that native speakers of English or Dutch do 80 percent of their talking with no more than about 500 different syllables18 (although these languages have many more than 10,000 different syllables). The syllable is posited as a repository of such overused, high-frequency syllabic gestures, one ‘syllabic score’ for each. Each time a new phonological syllable, such as /s/ or /sk/, or /t/, is composed, the corresponding general score is triggered. The score specifies which motor tasks (such as closing the glottis or releasing lip closure) are to be performed19 in order to generate the syllable. In WEAVER there is always competition among general scores. The activation spreads from individual segments to all syllable scores in which they participate (see Fig. 2). Hence, similar syllabic scores tend to be co-activated. The occasional mis-selection will resemble the target gesture. Selection latency is determined by Luce’s rule (as in the case for lemma selection).

There are further restrictions in selecting a syllabic score for execution. Repeat use of a particular type of syllable, for instance in producing the nonsense phrase hom-o-elf (where hom and the following elf are both CV syllables), may facilitate articulation20. General scores of similar types (such as CV or CVC) can apparently co-activate one another. Finally, WEAVER and the two-step interactive model have a featural representation of each segment. In both models the units of phonological encoding are whole phenomena (for which there is good experimental evidence21), but their features, such as ‘voiced’, ‘nasal’, ‘sonorant’, are already ‘stable’ in the process of syllabification (see legend to Fig. 3). During the next stage, phonetic encoding, these features function in the construction of articulatory gestures. The study of speech movement planning has become a discipline of its own22,23 and is not covered in the present review.

Conclusion

There is still a long way to go before the two research traditions emerging from speech error analysis and from naming chronometry are fully reconciled. But there has been lively and highly constructive interaction, leading to a much improved understanding of the processes involved in lexical selection and phonological encoding. One unifying force has been computational modeling. Current implemented models share their major traits; they are localist and symbolic; they compute quite similar linguistic representations. Another unifying force will hopefully proceed from brain imaging (see Ref. 57 for a recent review of imaging studies of word processing). It is the processing models that should guide the design of brain imaging experiments in word production, not native intuition as is too often the case24. The return will be convergence of evidence for or against particular processing components and their interactions.

Acknowledgement

I gratefully acknowledge helpful commentary by Antje Meyer and by Gary Dell.

References

3 Levelt, W.J.M. (1986) Speaking from Intention to Articulation, MIT Press
7 Dell, G.S. (1986) A spreading activation theory of retrieval in sentence production Psychol. Rev. 93, 382–439
10 Dell, G.S. et al. (1986) Lexical access in aphasic and non-aphasic speech. Psychol. Rev. 93, 801–837
18 Rondal, E. et al. (1983) Basic objects in natural categories Cognit. Psychol. 13, 318–319

language acquisition Cognition 43, 1–37
23 Kawamura, N. and Nishihara, Y. (1982) From lexical concepts to lexical
items Cognition 43, 23–42
word retrieval J. Psycholinguist. Res. 26, 33–67
25 Martin, N. et al. (1990) Phonological facilitation of semantic errors in
26 Brown, J.S. (1988) Grosberg and colleagues asked the hypoten-
27 Glaser, M.D. and Dinges, F.H. (1986) The time course of
picture-word interference J. Exp. Psychol. Hum. Percept. Perform. 7,
1267–1277
speaking Cognition 42, 157–183
influences on visual object naming errors in optic aphasia: a
speaking: some uptake to psychology. In W. van Millingen (Ed.) 269, 517–526
33 Levelt, W.J.M. et al. (1991) The time course of lexical access in speech
production: a study of picture naming Psychol. Rev. 98, 132–167
34 Dell, G.S. and O'Seaghdha, P.G. (1991) Mediated and convergent
lexical priming in language production: a comment on Levelt et al.
Psychol. Rev. 98, 804–814
prosody in lexical access in speech production: further evidence from
the coactivation of non-phonemic J. Exp. Psychol. Learn. Mem. Cognit. 12, 577–581
objects J. Exp. Psychol. 17, 273–281
Neurosci. 10, 553–567
speech production: retrieval of syntactic information and of
gender is on the tip of Italian tongues Psychol. Sci. 8, 214–217
41 Levelt, W.J.M., Mirus, M. and Zanetti, B. (1986) The two-stage model of
lexical retrieval: evidence from a case of agrammatism with selective preservation of grammatical gender Cognition 17, 190–216
and phonological knowledge in lexical access: evidence from the ‘tip–
of-the-tongue’ phenomenon Cognition 49, 369–384
lemma/knower distinction in models of speech planning comment on
44 Vigliocco, G. et al. ‘I count and I can’t find’ information available when the
noun is not: an investigation of tip of the tongue states and anoma
J. Mem. Lang. (in press)
45 King, T. (1992) Prelexical and postlexical features in language
production Appl. Psycholinguist. 13, 199–220
46 Mayo, E. Gender processing in speech production: evidence from
German–English bilinguals. In (in press)
speech error analysis: achievements, limitations, and alternatives
Cognition 42, 161–211
the production of spoken words J. Exp. Psychol. Learn. Mem. Cognit. 24, 630–656
50 Indefrey, P. and Levelt, W.J.M. in The Cognition Neuroscience (2nd
edn) (Ed.) (in press)
51 Levelt, W.J.M. and Hoedssen, L. (1992) Do speakers have access to a
mental syllable? Cognition 45, 219–260
52 Breman, C.P. and Goldstein, L. (1998) Some notes on syllable
structure in articoxiality phonology Psychol. Rev. 48, 150–155
54 Roelofs, A. (1988) Phonological segments and features as planning
production, in Principles of Experimental Phonology (Lane, N.L., ed.)
pp. 3–45, Oxford University Press
investigation of reference frames for the planning of speech
movements Psychol. Rev. 102, 411–433
57 Price, C.J. (1998) The functional anatomy of word comprehension and
production Trends Cognit. Sci. 2, 281–287

Coming soon to
Trends in Cognitive Sciences

• Multistable phenomena – changing views in perception, by D.A. Leopold
and N.K. Logothethis
• Possible stages in the evolution of language capacity, by R. Jackendoff
• Motion transparency: making models of motion perception transparent, by R.J.
Snowdon and A.J. Venraten
• Lessons from children with specific language impairment, by J.B. Tomblin and J. Pandich
• Reply from H. van der Lely
• Multiple determinants of image segregation, by M.A. Peterson
• Speech segmentation and word discovery: a computational perspective, by M.R. Brent
• Visual perception of self-motion, by M. Lappe, F. Bremmer and A.V. van den Berg
• Possible stages in the evolution of language capacity, by R. Jackendoff
• Motion transparency: making models of motion perception transparent, by R.J.
Snowdon and A.J. Venraten
• Lessons from children with specific language impairment, by J.B. Tomblin and J. Pandich
• Reply from H. van der Lely
• Multiple determinants of image segregation, by M.A. Peterson
• Speech segmentation and word discovery: a computational perspective, by M.R. Brent
• Visual perception of self-motion, by M. Lappe, F. Bremmer and A.V. van den Berg

Corrigendum


On p. 363, it was incorrectly stated that ‘Children with attention deficit hyperactivity disorder have also been shown to have statisti-
cally smaller vermal lobules VI and VII on MRI (Ref. 18), and a similar observation has been made in fragile-X syndrome.’

We apologize to readers for this error.

Ph: 51364-6119/(99)01326-1