Models of word production

Willem J.M. Levelt

Research on spoken word production has been approached from two angles. In one research tradition, the analysis of spontaneous or induced speech errors led to models that can account for speech error distributions. In another tradition, the measurement of picture naming latencies led to chronometric models accounting for distributions of reaction times in word production. Both kinds of models are, however, dealing with the same underlying processes: (1) the speaker’s selection of a word that is semantically and syntactically appropriate; (2) the retrieval of the word’s phonological properties; (3) the rapid syllabification of the word in context; and (4) the preparation of the corresponding articulatory gestures. Models of both traditions explain these processes in terms of activation spreading through a localist, symbolic network. By and large, they share the main levels of representation: conceptual/semantic, syntactic, phonological and phonetic. They differ in various details, such as the amount of cascading and feedback in the network. These research traditions have begun to merge in recent years, leading to highly constructive experimentation. Currently, they are like two similar knives honing each other. A single pair of scissors is in the making.

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 person. 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 words. 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 it is 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 production. 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 exception, 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 phonemic studies of naming.

The speech error tradition

In the late 19th century, psycholinguists began to publish a substantial corpus of German speech error data that they had diligently collected (Ref. a). The corpus, along with the theoretical analyses they provided, are still relevant today as a foundation for speech error rules. One important distinction they made was between meaning-based substitutions (such as Iber (‘horse’) for motre (‘may’)) and form-based substitutions (such as stude (‘student’) for studter (‘teacher’)), acknowledging that there is often a phonological connection in meaning-based errors (i.e. the over-representation of mixed errors was observed over a century ago). Freud was quick to confuse the now generally accepted distinction between meaning- and form-based errors by claiming that innucent form errors are practically all meaning-driven (why does a patient say of her parents that they have Guiz (‘good’) instead of Guiz (‘clevermen’)? 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 (weil [‘weil’] for well [‘well’]), anticipations (though [‘though’] for further [‘further’]), perseverations (he not okay for he nor okay) and blends or contaminations (seemed, blending seem and end).

Many linguists and psychologists have continued this tradition (Ref. b), but an elegant consensus (probably triggered by the work of Cohen, Ref. c) began in the late 1960s. In 1973, Fromkin edited an influential volume of speech error studies, with part of her own collection of errors as an appendix (Ref. d). Another substantial corpus was built up during the 1970s, the MIT-CU corpus, led to two of the most influential models of speech production: (1) Garrett discovered that word exchanges (such as he left it and forget it behind) can span some distance and mostly preserve grammatical category as well as grammatical function within their clauses (Ref. e). Sound/form exchanges (such as nail pit for nail pit) on the other hand, ignore grammatical category and preferably happen between closely related words. These indicate 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 (morpheme, phoneme) is organized. (2) Shattuck-Hufnagel’s scan-copier model connects phonological encoding (Ref. f). A core notion here is the existence of phonological frames, which are automatically activated for words that are orthographically (not semantically) related to the word itself. The main finding was that color naming is substantially slowed down when the colored word is a different color name. It is, for instance, difficult to name the word green when it is written in red. But naming the word was not affected by the word’s color.

Rosinski et al., interested in the automatic word reading skills of children, transformed the Stroop task into a picture/word interference task (Ref. g). The children named the word of the corresponding printed object. This started a research tradition of measuring naming latencies, naming objects and words. The main finding was that color naming is substantially slowed down when the colored word is a different color name. It is, for instance, difficult to name the word green when it is written in red. But naming the word was not affected by the word’s color. The dominant current view is that there is a direct access route from the word to its phonological code, whereas the line drawing first activates the object concept, which in turn causes the activation of the phonological code – an extra step. Another classical discovery in the picture-naming tradition (by Oldfield and Wingfield; Ref. h) is the word frequency effect (see main article).

In 1953, Stroop introduced a new research paradigm, now called the ‘Stroop task’ (Ref. i). The stimuli are differently colored words. The subject’s task is either to name the color or to say the word. Stroop studied what happened if the word was a color name itself. The main finding was that color naming is substantially slowed down when the colored word is a different color name. It is, for instance, difficult to name the word green when it is written in red. But naming the word was not affected by the word’s color. Rosinski et al., interested in the automatic word reading skills of children, transformed the Stroop task into a picture/word interference task (Ref. g). The children named the word of the corresponding printed object. This started a research tradition of measuring naming latencies, naming objects and words. The main finding was that color naming is substantially slowed down when the colored word is a different color name. It is, for instance, difficult to name the word green when it is written in red. But naming the word was not affected by the word’s color. The dominant current view is that there is a direct access route from the word to its phonological code, whereas the line drawing first activates the object concept, which in turn causes the activation of the phonological code – an extra step. Another classical discovery in the picture-naming tradition (by Oldfield and Wingfield; Ref. h) is the word frequency effect (see main article).

The chronometric tradition

In 1895, Carmeli (Ref. j) discovered that naming a list of 100 line drawings of objects took twice as long as naming a list of the corresponding printed object names. This started a research tradition of measuring naming latencies. It could not be attributed to practice. It could not be attributed to the differences between line drawings and words. Fraisse showed that when a small circle was named as ‘circle’, took on average, 619 ms, but when named as ‘ring’ took 158 ms. Fraisse concluded that the task induced different codes to be accessed. They are not graphemic codes, because Potter et al. obtained the same picture-word difference in Chinese (Ref. k). The dominant current view is that there is a direct access route from the word to its phonological code, whereas the line drawing first activates the object concept, which in turn causes the activation of the phonological code – an extra step. Another classical discovery in the picture-naming tradition (by Oldfield and Wingfield; Ref. h) is the word frequency effect (see main article).

The corpus, along with the theoretical analyses they provided, established a general distinction between meaning-based and form-based errors. This distinction was later formalized by the WEAVER model (Ref. l), which provided a framework for understanding the differences between line drawings and words. Fraisse showed that when a small circle was named as ‘circle’, took on average, 619 ms, but when named as ‘ring’ took 158 ms. Fraisse concluded that the task induced different codes to be accessed. They are not graphemic codes, because Potter et al. obtained the same picture-word difference in Chinese (Ref. k). The dominant current view is that there is a direct access route from the word to its phonological code, whereas the line drawing first activates the object concept, which in turn causes the activation of the phonological code – an extra step. Another classical discovery in the picture-naming tradition (by Oldfield and Wingfield; Ref. h) is the word frequency effect (see main article).
nodes represent whole linguistic units, such as semantic features, syllables or phonological segments. Hence, they are all ‘symbolic’ models. Of the many models with ancestry in the speech error tradition only a few have been computer-implemented. Among them, Dell’s two-step interactive activation model has become by far the most influential. Figure 1 represents a fragment of the proposed lexical network.

The network is called ‘two-step’, because there are two steps from the semantic to the phonological level. Semantic frame nodes spread their activation to the corresponding word or lemma nodes, which in turn spread their activation to phoneme nodes. Activation ‘cascades’ from level to level over all available connections in the network. The type of model is called ‘interactive’, because all connections are bi-directional; activation spreads both ways. Interactiveness is a property shared by all models in this class. One of the original motivations for implementing this feature is the statistical over-representation of so-called mixed errors in speech errors corpora. They are errors that are both semantic and phonological in character. If, for example, your target word is cat but you accidentally produce rat, you have made a mixed error. The network in Fig. 1 can produce that error in the following way. The lemma node cat is strongly activated by its characteristic feature set. In turn, it spreads its activation to its phoneme nodes /k/, /æ/ and /t/. A few of the semantic features of cat (such as ‘animal’ and ‘mammal’) co-activate the lemma node of rat. But the same lemma node rat is further activated by feedback from the new active phonemes /æ/ and /t/. This confluence of activation gives rat a better chance to emerge as an error than either the just semantically related dog or the just phonologically related mat. Interactiveness also gives a natural account of the tendency for speech errors to be real words (for example mat rather than get). Still, bi-directionality needs independent motivation (in functionality it can hardly be to induce speech errors). One recurring suggestion in this class of models is that the network_error activates itself in both word production and word perception. That would, of course, require bi-directionality of the connectivity. However, Dell et al. argue against this solution because many aphasic patients show both good auditory word recognition and disturbed phonological encoding. The functionality of bi-directional connections (and hence interactiveness) would rather be to support fluency in lemma selection. Some word forms, in particular the ones that are infrequently used, are less accessible than others. It will be advantageous to select a lemma whose phonological form will be easy to find. Feedback from the word form level will provide that function. (And might explain a recent chromatic effect). Still, one should consider the possibility that interactiveness is merely a property of the error mechanism: an error might occur precisely then when undue interactivity arises in an otherwise discrete system.

Most implemented computational models in the chronometric tradition extend no further than accessing the word’s whole name from a semantic or conceptual base. There is no activation of phonological segments, no phonological encoding. Only Roesler’s WEAVER model has a fully developed phonological component. A fragment of the WEAVER lexical network is shown in Fig. 2.
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\(^1\), 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 \textit{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 \textit{rat} for \textit{cat}) will have a better chance of ‘slipping through’ the monitor than one that is semantically totally out of place (such as \textit{mat} for \textit{rat}). 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\(^1\). 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 \textit{cat} or \textit{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 \textit{cat}, \textit{animal}, \textit{siamese} or anything else. Rosch\(^2\) has shown that we prefer ‘basic level’ terms to refer to objects (\textit{cat} rather than \textit{animal}; \textit{dog} rather than \textit{collie}, etc.), but the choice is ultimately dependent on the perspective you decide to take on the referent for your interlocutor\(^3\). Will it be more effective for you to refer to my sister as \textit{my sister} or as \textit{that lady} or as \textit{the physicist}? It will all depend on shared knowledge and discourse context. This freedom of perspective-taking appears quite early in life\(^4\) 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.

Fig. 2. Fragment of Roelofs’s WEAVER network model of the lexicon. The nodes in the upper layer represent whole lexical concepts. The arrow connections represent the semantic relations holding among them. There is bi-directional activation spreading at this level. The mid stratum is syntactic. The nodes represent lemmas, that is syntactic words and their features. Among these properties are the word’s syntactic frame, its variable inflectional features (such as number), its gender (in gender marking languages), etc. The activation spreading is uni-directional from lemma node to feature nodes. The lemma’s connection to its concept node represents the sense of the word. The connection allows for bi-directional activation spreading. The arrow down from the lemma points to its form. Activation spreading is uni-directional here, which is maintained all the way down through the network. Only a selected lemma can spread its activation to the form level. The bottom stratum represents morpheme nodes with their connections to metrical and phoneme nodes. In their turn, the phoneme nodes point to all (stored) phonetic syllables in which they participate; they are not specified for their syllable position. There are no inhibitory connections in the network. (Adapted from Levelt.)
Step 1. Accessing the morpho-phonological code

Step 2. Spelling out the phonological code

Step 3. Prosodification

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.

Lexical selection

In the chronometric tradition lexical selection has been studied with interference paradigms, in particular pictures-word interference (see Box 1). The recurring finding has been that naming an object is slowed down when a distractor word is presented with the picture; the effect is stronger when the distractor word is semantically related to the target than when it is semantically unrelated and it is at maximum when picture and distractor word are presented simultaneously.

The WEAKER model provides an accurate quantitative account of a wide range of pictures-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 lemma. In the semantic network activation spreads to related concepts, such as goat and llama. They, in turn, spread 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 Lucie’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 WEAKER 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 stronger 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) 

There is a substantial literature on the types of semantic (and other) errors produced by aphasic patients, which will not be covered in the present review. It is a major challenge to predict these error distributions by ‘damaging’ the normal network. Dell et al. have set an impressive example. 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.

The timing of lexical selection is not explicitly modeled in the speech-error based models. In the interactive two-step model the selection moment is determined from outside. 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,c). 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, local and time; 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 ‘s’. 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 intermingling the A-B pairs in such a way that they no longer share their first syllable. For instance, the first set of the homogeneous condition (fourth column in the table) has lower, local and major as response words. Each word pair is its own control in the experiment: it appears both in the homogeneous and the heterogeneous condition.

In the homogeneous 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.

Table 1. The implicit priming method: priming the first syllable of bisyllabic words

<table>
<thead>
<tr>
<th>Homogeneous condition</th>
<th>Heterogeneous condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set 1</strong></td>
<td><strong>Set 2</strong></td>
</tr>
<tr>
<td>single-toner</td>
<td>signal-beacon</td>
</tr>
<tr>
<td>place-local</td>
<td>priest-beadle</td>
</tr>
<tr>
<td>fruit-lotus</td>
<td>glass-beaker</td>
</tr>
<tr>
<td><strong>Set 1</strong></td>
<td><strong>Set 2</strong></td>
</tr>
<tr>
<td>single-toner</td>
<td>place-local</td>
</tr>
<tr>
<td>signal-beacon</td>
<td>priest-beadle</td>
</tr>
<tr>
<td>captain-major</td>
<td>cards-maker</td>
</tr>
</tbody>
</table>

Morpho-phonological encoding

When you are planning the sentence ‘they are selecting me’, you must retrieve from your long-term memory the morpho-phonological codes for each of the selected words, among them the two morpheme-size codes select and ing (see Fig. 5), and compute their syllabification and accent structure in context (so-de-ting). This naturally divides the process into ‘code retrieval’ and ‘syllabification’.

Code retrieval

An item’s morpho-phonological code consists of its morphological makeup, its metrical shape and its segmental makeup (see Fig. 3, Step 1 and Step 2). Retrieving that information follows activation/seletion of the lemma3. Each ink and many subjects have been spilled over this issue. In the WEAVER model, the activation and retrieval of a morpho-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 morphological code (kat). This predicts that alternative active, but non-selected lemmas (such as the lemma for age) do not spread any activation to their morphological code. Initial experimental evidence3 showed that, in picture naming, there is semantic but indeed no morphological activation of same-category alternatives (if cat is the target, dog is semantically but not phonologically active).
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 extent. However, the original finding was reconfirmed in a quite critical replication. Still, evidence for phonological co-activation of semantic alternatives was obtained for one restricted case: if the alternative is a synonym of the target. When you name the picture of a couch, the phonological code of a couch 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 lemmas 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 clare), hardly ever of same-category terms.

There is a strong and robust word-frequency effect in word production (which is in part an age-of-acquisition effect). Controlling for conceptual biases, you are typically faster in producing a high-frequency response such as word than a low-frequency response such as word. It is now known that the effect arises in accessing the phonological code (Fig. 3, Step 3), not in selecting the lemma. 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 word and this also holds for the much amplified case of word finding trouble in many amnestic patients. This has been used as one of many arguments for the distinction between an ‘earlier’ semantic lemma-level and a ‘later’ phonological code level in the lexical network. But that argument has provoked some controversy—which is, so far, unresolved. Probably more relevant speech error evidence for the precedence-of-semantic claim is the repeated finding of almost absolute gender preservation in phonological word substitution errors (such as deixa for dezigna in Italian). 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, t) and (l, η). Syllabification proceeds ‘from left to right’. You first chunk the first two phonemes to create the syllable /sl/. You then take the next three to compose the syllable /kt/ and, finally, you chunk the remaining segments to compose /t/ /l/. 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, /t/ /l/, straddles two morphemes, select and ing. This can also happen across words. When you utter They will select us, the syllabification will be /si-lkt-t/ /ə/, where /t/ /k/ straddles the words select and us. But when you produce they select me, the syllabification is /si-lk/ /ə/, without straddling. Apparently, the syllables are not given in the phonological code of the morpheme, but depend 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; they in select will appear as syllabic onset or as syllabic offset, dependent on the context. The domain of syllabification (such as selecting, select, 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 varies among languages. These rules are rapidly applied, time 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, CVc, 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 convincing. In particular, the idea that phonemes in the phonological code (such as /t/ in select) are marked for a particular syllable position creates more problems than it solves. The preference of sound exchanges (such as maggy kenvia for maggy marvia) to preserve syllabic position can be explained differently, as a combination of word onset vulnerability, phoneme similarity and phonotactic restrictions.

There is good phonetic evidence, however, for the existence of moratal frames (see Fig. 3). For Dutch, and probably for other stress-assigning languages such as English and German, there is a dominant moratal 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? How is 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 hardly faster than naming a long word?
- Which brain regions subserve the core components of conceptual/semantific preparation, lexical selection, phonological code retrieval, prosodification, phonetic encoding, articulation and self-monitoring?
word stress goes to the first full-vowelled syllable (morning, yellow, forget – the ‘s’ in the latter word is not full-vowelled, but rather a neutral ‘sub-voiced’ sound). This can be automatically produced in incremental syllabification. But when a word has a deviant stress pattern, the automaticity breaks down10,11 (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 production12 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 shapes. 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 score. Statistics show that native speakers of English or Dutch do 80 percent of their talking with no more than about 500 different syllables13 (although these languages have many more than 10 000 different syllables). The syllable is posited as a repository of such oversimplified, high-frequency syllabic gestures, one ‘syllabic score’ for each. Each time a new phonological syllable, such as /bi/, /ki/, or /t/ (i.e., is composed, the corresponding gestural score is triggered. The score specifies which motor tasks (such as closing the glottis or releasing lip closure) are to be performed14 in order to generate the syllable. In WEAVER there is always competition among gestural scores. The activation spreads from individual segments to all syllabic 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 was the case for lemma selection). There are further restrictions in selecting a syllabic score for execution. Repeated use of a particular type of syllable, for instance in producing the nonsense phrase home-oil-feer (where feer and the following syllable are both CVC syllables) may facilitate articulation15. Comical scores of similar types (such as CV or CYC) 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 phonemes (for which there is good experimental evidence16), 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 own17,18 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 in the processing models that should guide the design of brain imaging experiments in word production, not naive intuitions as is too often the case19. 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


