Chapter 2
Properties of Spoken Language Production

Zeni M. Griffin and Victor S. Ferreira

Although a common caricature of speaking is that it is the reverse of listening, language production processes fundamentally differ from comprehension processes in many respects. Whereas people typically recognize the words in their native language quickly and automatically, the same words require an intention to speak and can take over five times longer to generate than to recognize. For example, listeners begin to direct their gaze to the referent of a spoken noun (even in the absence of highly predictable speech) before the speaker completes articulation of the word (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), whereas speakers typically take about 900 ms to begin to generate a noun in isolation based on a pictured object (e.g., Snodgrass & Yuditsky, 1996).

Language production is logically divided into three major steps: deciding what to express (conceptualization), determining how to express it (formulation), and expressing it (articulation; Levelt, 1989). Although achieving goals in conversation, structuring narratives, and modulating the ebb and flow of dialogue are inherently important to understanding how people speak (for review, see Clark, 1996), psycholinguistic studies of language production have primarily focused on the formulation of single, isolated utterances. An utterance consists of one or more words, spoken together under a single intonational contour or expressing a single idea (e.g., Boomer, 1978; Ferreira, 1993). While Ferreira and Englehart’s chapter in this volume on syntax describes processes that allow speakers to produce their words in grammatical utterances, we focus instead on the processing of the words themselves. Indeed, most theories of multi-word utterance or sentence production ultimately boil down to an account of how sentences acquire their word orders and structures, how the dependencies between words are accommodated (e.g., subject–verb agreement), and a functionally independent account of how individual content words are generated (e.g., Chang, Dell, Bock, & Griffin, 2000; Ferreira, Kempen & Hoenkamp, 1987). In this chapter, we describe the basic properties of spoken word production, outlining empirical data that demonstrate the properties of the processes resulting in speech and discussing the processing assumptions that models of language production invoke to account for these properties. Although it could easily fill a chapter of its own, we conclude by discussing timing in multi-word utterances.
1. GENERATING WORDS

The simplest meaningful utterance consists of a single word. Generating a word begins with specifying its semantic and pragmatic properties. That is, a speaker decides upon an intention or some content to express (e.g., a desired outcome or an observation) and encodes the situational constraints on how the content may be expressed (e.g., polite or informal speech, monolingual or mixing languages; see Levelt, 1989). This process, termed conceptualization or message planning, is traditionally considered prelinguistic and language neutral (Garrett, 1975; Levelt, 1989). However, speakers may include different information in their messages when preparing to speak different languages (see Slobin, 1996, on thinking for speaking). The next major stage is formulation, which in turn is divided into a word selection stage and a sound processing stage (Fromkin, 1971; Garrett, 1975). Deciding which word to use involves selecting a word in one’s vocabulary based on its correspondence to semantic and pragmatic specifications. The relevant word representation is often called a lemma (Kempen & Huijbers, 1983), lexical entry, lexical representation, or simply a word, and it marks the presence of a word in a speaker’s vocabulary that is capable of expressing particular semantic and pragmatic content within a particular syntactic context. Sound processing, in contrast, involves constructing the phonological form of a selected word by retrieving its individual sounds and organizing them into stressed and unstressed syllables (phonological encoding) and then specifying the motor programs to realize those syllables (phonetic encoding). The final process is articulation, that is, the execution of motor programs to pronounce the sounds of a word. This sequence of stages is illustrated in Figure 1.

This gross analysis of language production serves to illustrate the complexity of expressing an idea in words. The challenges posed by this complex problem give rise to the fundamental properties of the word production process, the descriptions of which below form the bulk of this chapter. Roughly, these properties delineate steps in processing (Property 1), describe how speakers deal with the relationship between meaning and word (Properties 2–9), explain how speakers represent and assemble the sounds of words (Properties 10–13), and how these processes play out in time (Properties 14 and 15).

Interestingly, current models of word production agree on the basic facts about how the system works to a surprising extent, with only minor variations in explanatory mechanisms. When models differ, the tendency is to focus on different stages of production, such as word selection or phonological encoding, and different aspects of these stages such as speed of processing or how processing may go awry to yield speech errors. This means that if the properties of production described below, most are accounted for (at least to some level of detail) by most models of production. Other properties of production have yet to receive detailed attention, although we feel that much can be gained if these properties are addressed in future theories. Next, we describe each of these properties of word production in turn.

1.1. Basic Steps of Word Production

1.1.1. Property 1: Word selection precedes sound assembly

When speakers access word representations, they do so first based on meaning and then focus on assembling their sounds. Several sorts of evidence suggest this. The first and strongest evidence comes from analyses of errors made during spontaneous speech (Fromkin, 1971; Garrett, 1975), which reveal that speech errors most frequently involve units that can be most conservatively considered to correspond to whole words, morphemes (i.e., minimal units of meaning such as can and berry in cranberry), or individual speech sounds (i.e., phonemes or segments such as the b- and oo-sounds in boot). In particular, error patterns suggest that a speaker may err in selecting a word but correctly assemble and pronounce its component sounds, or they may successfully select a word that can express an intended meaning, but then err in assembling its sounds. Table 1 lists examples of word, morpheme, and sound errors. In addition, the word production process occasionally falters at a point where speakers seem to have selected a word to express what they want to say but have not yet retrieved all of its sounds (see Property 8 for details).

Another sort of evidence that is often cited as showing that meaning-based word selection precedes the processing of sounds’ sounds comes from experiments exploring the time course of word production (Schieffers, Meyer, & Levelt, 1990). Specifically, when speakers name pictures as they ignore distractor words, semantically related distractors (e.g., reading tiger when naming a picture of a lion) primarily slow object-naming latencies when the distractor word appears simultaneously with the object or precedes it by up to 400 ms (Glas & Døngelhoff, 1984; Roelofs, 1992; Starreveld & La Heij, 1995, 1996; note that the effects of semantically and phonologically related words on word production...
are the reverse of what is typically observed for primes in word recognition – for review, see Balota, Cortese, & Yap, this volume). When a semantically related word is presented after the object appears, it either has no impact on latencies or results in faster latencies relative to an unrelated control word. In word translation (which takes a bit longer than object naming), a semantically related distractor word speeds production when it precedes the to-be-translated word by 400 ms and slows it when presented 200 ms after the to-be-translated word (Bloem & La Heij, 2003). In contrast, phonologically related distractor words (e.g., hearing fat when naming a picture of a lion) sometimes have no effect on object naming latencies when they precede the presentation of the objects, but consistently facilitate naming latencies when they appear at the same time or after the objects (Damin & Martin, 1999; Schriefers et al., 1990). That said, phonologically related distractors presented as early as 300 ms before objects may facilitate naming (Starreveld, 2000). Interpretation of effects in the picture-word interference paradigm relies heavily on uncertain assumptions about word recognition and how it interacts with word production (of Levitt, Roelefs, & Meyer, 1999; Roelefs, Meyer, & Levitt, 1996; Starreveld, 2000; Starreveld & La Heij, 1995).

That speakers produce words first by processing their meaning-level properties, then by processing their sound-level properties is arguably the fundamental property of word production. Models of word production incorporate this property by assuming two stages in producing words as well as separate word-level and sound-level representations (e.g., Caramazza, 1997; Dell, 1986; Levitt et al., 1999, but see Starreveld & La Heij, 1996, for an exception). Meaning-level representations make lexical-level representations available, which in turn provide access to individual sound-segment representations. This implies that it is not possible to go from meaning to sounds except through a lexical representation.

### 1.2. Selecting A Content Word

Despite the apparent simplicity of the resulting utterance, production of a single word can go awry in a number of ways and can take a surprisingly long time under some circumstances. Studies of isolated word production have focused primarily on nouns (e.g., person, place, or thing) with some studies of verbs (i.e., action words and predicates), ignoring other grammatical classes of content words that are less often spoken alone. In one-word utterances, the properties of word production processes appear similar for nouns and verbs (e.g., MacKay, Connor, Albert, & Obler, 2002; Vigliocco, Vinson, Damin, & Levitt, 2002). There is no reason to suspect that other types of content words are prepared differently in single word production.

#### 1.2.1. Property 2: The intention to produce a word activates a family of meaning-related words

Speech error analyses suggest that the most common error in word selection occurs when a speaker substitutes a semantically related word for the intended one, such as calling a van bus (Dell et al., 1997). A related type of speech error is a blend in which two words that could sensibly fill a particular slot in an utterance are spliced together to form an unintended string of sounds, such as behavior and derartment emerging as behorphment (Wells, 1951/1973). Often the words that form a blend are not true synonyms out of context but are interchangeable only within the context of the utterance (Garrett, 1975).

Such observations suggest that accessing word representations by meaning or message representations is not surgical. Specifically, the intent to produce a particular word will lead to the activation of a family of words, all sharing some aspect of the intended meaning. This leads to the issue of how meaning is represented in models of word production, and on this issue, two major theoretical positions have been staked. On one side are decompositional views of semantic representation (Bierwisch & Schreuder, 1992; Katz & Fodor, 1963). Decompositional views portray the primitives of semantic representation as being entities that are smaller than the words whose production they ultimately support.
For ease of exposition, such features are sometimes described as themselves meaningful, so that the meaning of bird might include HASE WINGS, HAS FEATHERS, SINGS SONGS, and the like (e.g., Certe & McCrae, 2003; Vigliocco, Viswan, Lewis, & Garrett, 2004). However, decomposition approaches are equally (and perhaps more realistically) assume that lexicalizable concepts consist of organized collections of arbitrary features (bearing a non-trivial resemblance to parallel-distributed processing accounts of cognition; McClelland & Rumelhart, 1986; Rogers & McClelland, 2004) rather than semantic primitives (or other tidy features). By either account, semantic similarity arises straightforwardly from feature overlap—lexical items are similar to one another to the extent that the semantic features that promote their use are the same. In turn, the activation of a family of words, all related in meaning, also follows straightforwardly—if to produce bird, a speaker must activate the features HASE WINGS, HAS FEATHERS, and SINGS SONGS, then other words that also share those features will become activated (e.g., airplanes, because they have wings, opera singers, because they sing songs), and to an extent that increases with the number of shared features. Furthermore, the observation that many meaning-level influences on lexical production are context-dependent (e.g., semantic word substitutions and blends typically only involve words that can be interchanged within a particular context; Garrett, 1975; Levelt, 1988) follows from the natural assumption that representations of meaning are activated in context-dependent fashion. Only words of the same level of specificity interfere with one another in the picture-word interference task, which also suggests that specificity is an important feature or constraint on word activation (Costa, Mahon, Savova, & Caramazza, 2003; Vlakovitch & Tyrell, 1999). Theories of production that posit decomposition of semantic features include Ogood (1963), Fosskin (1971), Del (1980), Stemperger (1985), Butterworm (1989), Caramazza (1997), and Chang et al. (2003). In addition, decomposition theories have played an important role in the development of connectionist models of word processing (e.g., Strain, Patterson, & Seidenberg, 1995) and language deficits (e.g., Hinton & Shaft, 1991).

On the other side are non-decomposition views of semantic representation. The philosophical case for non-decompositional views has been forwarded most prominently by Fodor (1975). With respect to word production, the WEAVER++ model (Levelt, 1989, 1992; Roes, 1993) and other models (e.g., Blean & La Heij, 2003; Garrett, 1982; Stare & van & La Heij, 1996) have adopted non-decompositional representations. According to non-decompositional views, the representational base of words and their meanings bear a one-to-one relationship, so that the word bird is led by an atomic meaning representation of BIRD, the word airplane is led by an atomic meaning representation for AIRPLANE, and so forth. These atomic meaning representations are often called lexical concepts. Within such accounts, the activation of a family of words, all similar in meaning, is not quite as straightforward as it is with decompositional accounts. Specific claims as to how multiple meanings become activated have been presented by Roes (1992) and Levelt et al. (1999). The idea is that activating the concept BIRD activates the concept FISH because FISH will be connected within a semantic network to the concept ANIMAL (through what is sometimes called an “is-a” link), which will then spread activation to BIRD. The concept BIRD can then activate the word fish. Other links through other mediating properties will cause other meaning-similar words to become active. Figure 2 illustrates these two different forms of semantic representations.

1 Words referring to words or lexical units are printed in lowercase italics, whereas words referring to semantic representations appear in uppercase italics. The meaning of a word is a semantic representation, although word is written in lowercase italics in such a context.

Figure 2. Diagrams based on WEAVER++ model (Levelt et al., 1999) on the left and models by Dell (e.g., Dell et al., 1997) on the right.
tendency to compete for selection (e.g., Cutting & Ferreira, 1999; Lupker, 1979). That is, competition is restricted to words that express similar meanings rather than simply related meanings.

According to nearly all models of word production, the availability or activation level of one word affects the speed and/or likelihood that a speaker will select another word. The simplest way of modeling this is to have the probability of selecting a word (or other unit) directly related to its level of activation, so that if an unintended word has too high a level of activation, it will be erroneously selected. Due to patterns of connectivity between semantically (and phonologically) related words, these are the most likely to be highly activated and selected in error. Several models use activation levels and connectivity alone to account for patterns of speech errors (e.g., Dell, 1986; Dell et al., 1997; MacKay, 1982). Although this type of selection mechanism suffices to account for error patterns, additional assumptions are needed to account for latencies.

A number of models explain differences in naming latencies with partial inhibition between activated word representations (e.g., Cutting & Ferreira, 1999; Harley, 1993; Stemberger, 1985). The more activated one word (or unit) is, the more it inhibits (decreases) the activation of other words. Words must reach some threshold before they are selected and the time it takes for one word to win out by suppressing others is reflected in naming latencies (for discussion of selection mechanisms in localist networks, see Schade & Berg, 1992). By contrast, in WEAVER++ (Lvell et al., 1999), the timing of word selection is influenced by two factors. The first is the activation of the to-be-selected word relative to all other activated lexical representations in a response set (embodiment is often termed a Luce ratio; Luce, 1959). As in many other models, the more activated a word is relative to other words, the more likely it is to be selected. The other factor is termed a critical difference, whereby a lexical representation can only be selected if its activation exceeds the activation of all other representations by some minimum amount. This critical difference is important, as it implies that an alternative representation with high activation might be a more formidable competitor than two alternative representations each with half of its activation (because the one with higher activation is more likely to exceed the critical difference threshold).

Word-production models also must account for why semantically related distractor words interfere more with word production than unrelated words do. Almost every model of word production that has aimed to explain semantic interference attributes it generally to the fact that a speaker tries to select the most appropriate word for a stimulus, the word representations of semantically related distractors are activated more strongly and so more readily compete for selection of the alternative form than semantically unrelated distractors (for one exception, see Costa et al., 2003). Generally, this is assumed to occur because the lexical representations of semantically related distractors receive activation from two sources: the distractor words themselves and through their semantic relation to intended words. The representations of unrelated distractor words, in contrast, do not receive this latter source of activation. For example, naming a picture of a lion, the lexical representation of tiger would be activated not only by the distractor word

---

1.2.3. Property 4: Competition for selection is constrained by grammatical class and contextual features

Not just any semantically related word competes for selection with the most appropriate word to express a meaning. Critically, the competition is limited to words of the same grammatical class. That is, only nouns substitute for nouns, verbs for verbs, and so on (Fromkin, 1971; Garrett, 1975; Nooteboom, 1983). Likewise, distractor words only appear to interfere with word production when they share grammatical class, verb transitivity (e.g., Schriefers, Teuerl, & Meinhausen, 1998), and other context-relevant syntactic features with the most appropriate or target word (Schriefers, 1993; Tabossi, Collina, & Sanz, 2002). Even when substituting words are only phonologically related to the word they replace, they show a strong tendency to share the same grammatical category (e.g., Fay & Cutler, 1977). Thus, constraints on maintaining the grammatical class of an intended word are stronger than the constraints on expressing the intended meaning.

Models of word production typically invoke different processing mechanisms to impose syntactic constraints on word selection as opposed to semantic constraints. Several models posit syntactic frames in which content words are inserted after selection. The selection mechanism is blind to word representations that do not fit the slot it is trying to fill, such that only a noun can fill a noun slot (e.g., Dell, 1986; MacKay, 1982; Stemberger, 1985).

1.2.4. Property 5: The speed and accuracy of selection is affected by specific meaning-level properties

A number of factors related to semantic representations and their mapping to word representations affect the speed and accuracy with which a word is selected and produced. The concreteness or imageability\(^2\) of a word is one such factor (Morris, Chappell, &
Ellis, 1997). Presumably, words with concrete, imageable meanings such as vampire, wind, and pine have richer semantic representations that guide word selection more efficiently than words with more abstract meanings such as fear, sense, and spirit (De Groot, 1992). Higher imageability or greater concreteness facilitates word translation (De Groot, 1989), generating associations (Cattell, 1889), and word recall (e.g., Martin, Lesch, & Baatha, 1999). However, highly imageable words appear more prone to substitution by semantically related words, perhaps due to a tendency to share more meaning features with other words (Martin, Saffran, & Dell, 1996). Similarly, names of objects from structurally similar categories and, in particular, living things seem more error-prone than artifacts (for comprehensive review and a theory, see Humphreys & Forde, 2001).

In addition, sentence context clearly influences the speed of word selection, probably through the influence of a combination of pragmatic, semantic, and syntactic constraints. The more predictable a word is in an utterance (based on other people's attempts to guess it from context), the less likely a speaker is to silently pause or say um before saying it in spontaneous speech (Goldman-Eisler, 1958b), in laboratory settings (Goldman-Eisler, 1958a), and the faster a speaker will label a corresponding object (Griffith & Bock, 1998).

Explaining such effects on word selection within the context of decompositional views is fairly straightforward. The properties that make particular word meanings imageable or concrete also bestow those meanings with additional features and thus richer meanings (e.g., Gordon & Dell, 2002, 2003; Hinton & Shallice, 1991). Likewise, sentence contexts may increase or sharpen the features specifying the meaning to be lexicalized. In turn, these additional features should increase the activation levels of consistent word representations while activating fewer potential competitors. With respect to imageability and concreteness, non-decompositional views might take an analogous stance, but rather than propose that imageable and concrete word meanings have richer meanings, they might propose that those meanings participate in more richly interconnected meaning networks, which might selectively promote the activation of imageable or concrete word-meanings.

1.2.5. Property 5: Attended objects do not necessarily lead to lexical activation

In the semantic interference effect, hearing or seeing a semantically related word interferes with generating another word. This suggests that word representations might be easily activated based on any strongly associated stimulus in the environment, even objects. For example, seeing a banana might activate the word banana to some extent, even in the absence of any intention to talk about it. Indeed, several models make this prediction (e.g., Humphreys & Forde, 2001; Roelofs, 1997). In contrast to the semantic interference effect that these theories would predict, viewing a semantically related object facilitates production relative to viewing an unrelated object (e.g., viewing a banana while trying to generate the word apple; Bloom & La Heij, 2003; Damian & Bowers, 2003). This suggests that word representations may only be activated by meaning or visual representations in the presence of an intention to communicate about them. Speakers even tend to gaze directly at objects while generating words to inaccurately label them (i.e., lie about them; Griffith & Oppenheimer, 2003). So, it seems that distractors easily influence production processing via language comprehension processes but not via object recognition processes (for some exceptions that may be due to failure to focus on what to express, see Harley, 1990).

However, the data regarding name activation for ignored visual information are not completely clear-cut. At the same time as there is evidence of perceived objects failing to show any semantic interference effects (Bloom & La Heij, 2003; Damian & Bowers, 2003), other researchers have found results that suggest phonologial information about ignored objects becomes available (Morsella & Miccozzi, 2002; Navarrete & Costa, 2005). Specifically, speakers are faster to name an object (e.g., a cat) in the presence of a distractor object with a phonologically similar name (cap) than an unrelated name (shoe). This is problematic given that all models require that phonological activation be mediated by word representations, so that under the same conditions that one sees phonological activation of names, it should be possible to detect semantic interference just as in other situations.

The possibility that speakers only linguistically activate words they intend to speak is an important characteristic for models of word production to take account of. The solution is to restrict activation from freely flowing from semantic representations to word representations, limiting the flow to meanings within a pre-verbal message (for examples, see Bloom & La Heij, 2003; Bloom, Van Den Boogaard, & La Heij, 2004). Specifically, Bloom and La Heij (2003) propose that until a semantic-level representation reaches a threshold level of activation, it is unable to influence word representations (only other semantic representations), and that an intention to speak is necessary to achieve that threshold activation. Note that this in essence introduces a kind of discreteness to the word-production process between the levels of meaning representation and linguistic representation.

Restricting activation flow is likely to have additional consequences and potentially explain other observations. For example, selective activation may resolve the seeming contradiction that on the one hand, imageable words are produced more quickly and accurately, whereas they are also relatively more prone to word substitution errors. Specifically, if imageable words have richer meanings or participate in richer semantic networks, then when accessed, they probably lead to the activation of a wider cohort of semantically related word meanings. When that widely activated cohort of word meanings does not extend to lexical activation, the most appropriate word representation should not suffer lexical-level competition, and so the greater activation of the intended word meaning should be to be easily selected. At the same time, if we assume that speakers sometimes select the wrong word meaning for production or fail to highly activate important features for distinguishing similar objects, then because of the wider semantic cohort, speakers should be more likely to select the wrong meaning to lexicalize or activate a wider range of competing words when generating highly imageable word meanings than less imageable meanings.

Restrictions on activation flow between semantic and word representations cannot be blindly added to any model of production. For instance, WEAVER++ requires that meaning-level representations freely activate lexical representations in order to explain...
lexical-competition effects. For example, if the concept FISH (as activated through “is-a” links from the concept BIRD) is unable to activate the lexical representation of fish because FISH is not in the message, lexical competition between fish and bird is not possible and the lexical competition effects described above go unexplained.

1.2.6. Property 7: Selecting words is subject to long-term repetition effects that resemble learning

Selecting a word to name an object or express a meaning has long-lasting effects on how easy it is to retrieve that word again to express a similar meaning. One manifestation of this is in repetition priming for naming objects that lasts over months (Cave, 1997) and retrieving the same name for different exemplars of the same type of object (e.g., multiple cars) over the course of an experimental session (Bartlett, 1974). This increase in availability is also reflected in perseveratory speech errors, such as calling a giraffe zebra after correctly naming a zebra (Viskovitch & Humphreys, 1991). Also relevant are observations of increased latencies to name objects or actions when presented with other items from the same taxonomic category (the semantic homogeneity effect; e.g., Vigliocco et al., 2002). Note that these semantic interference and strengthening effects only occur when speakers must select words to label pictures, as sentence completions, or in some other way that is based on meaning. The perseveratory effects are not produced by reading words aloud or categorizing words as names for artifacts or natural objects (e.g., Viskovitch & Humphreys, 1991).

Generally speaking, such long-term effects invite explanations in terms of learning. This highlights a notable gap in models of word production: Unlike the subfields of word reading (e.g., Seidenberg & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996), phonological assembly (Dell, Juliano, & Covington, 1993), and even sentence production (Chang, 2002; Chang, Dell, & Bock, 2006), no major model of word production has emerged that accounts for learning effects during word production. Long-term repetition priming effects and the semantic homogeneity effect could be explained via automatic strengthening of connections between meaning representations (most readily imagined as semantic features) and a selected word representation whenever the generated word successfully expresses the meaning. The fact that such long-term effects occur only with conceptually mediated production is consistent with an explanation that requires a mapping and selection rather than simple activation of meaning or word representations (as in categorization and reading). This implies that expressing the same meaning with the same word should subsequently be more efficient, but expressing a similar meaning with a different word should be more difficult. The landscape of models of word production would benefit if it included as implemented and fully developed model with learning principles that could express these kinds of effects.

1.2.7. Property 8: Word production can halt part of the way through the process

Slips of the tongue such as saying frish grotto for fish grotto (Fromkin, 1971) suggest that speakers may correctly select the words that they intend to say but then err in assembling their sounds. Sometimes, speakers fail to retrieve any of the sounds of a word that they want to say. Researchers refer to this as being in a tip-of-the-tongue orTOT state, from the expression “to have a word on the tip of one’s tongue.” TOT states occur most often for people’s names (Burke, MacKay, Worthley, & Wade, 1991), but can also be elicited in the laboratory using definitions of obscure words such as “What do you call a word that reads the same backwards as forwards?” TOTs are more likely for meanings with low imageability (Astel & Harley, 1996) and uncommon words (Burke et al., 1991; Harley & Bown, 1998). Some forms of brain damage cause people to experience similar word retrieval problems for common words. When trying to come up with one of these elusive words, both brain-damaged and unimpaired speakers are able to provide a great deal of general world knowledge associated with the word and, moreover, information about the word’s syntactic properties, such as whether it is a count or mass noun in English (Vigliocco, Martin, & Garrett, 1999), its grammatical gender in Italian, French, or other languages with fairly arbitrary grammatical categories for nouns (Badecker, Miozzo, & Zanuttini, 1995; Miozzo & Caramazza, 1997a), and identify the correct form of the auxiliary for sought-for verbs in Italian (Miozzo & Caramazza, 1997a). Although speakers cannot say the whole word they seek, they often can identify its first letter or sound, its number of syllables, and words that sound similar (Brown & McNeill, 1966). Providing with candidates for missing words, the speaker can (with exaggeration) reject unintended words and (with gratitude) identify intended ones. Diary studies indicate that in normal life outside the laboratory, speakers typically come up with the TOT word minutes or days later. In the lab, cueing the speaker with sounds from the missing word increases the likelihood that it will “spontaneously” occur to them (Jenga & Burke, 2000; Meyer & Bock, 1992) and providing a homophone before a TOT-eliciting stimulus makes a TOT less likely to occur (e.g., eliciting cherry pit makes speakers more likely to successfully name Brad Pitt; Burke, Locantore, Austin, & Chae, 2004).

Two models of production have sought to specifically explain TOT states: WEAVE++ (Lambrecht, 1999) and Node Structure Theory (Burke, 1987; Burke et al., 1991). TOT states starkly illustrate the architectural assumptions of WEAVE++. After having selected word representation to express a meaning, the retrieval of the next required representation, the lexeme or phonological form of the word, fails. The successful selection of the lemma representation explains speakers’ confidence that they know a word to express and their ability to report the word’s syntactic characteristics (which are stored at the same level in the theory), whereas the failed selection of the lexeme representation explains speakers’ inability to articulate the word. Node Structure Theory takes a similar stance, except without the lexeme (i.e., a complete lexical-phonological representation). Specifically, it postulates that a (non-phonological) lexical node fails to fully activate (allow selection of) phonological information (syllables, segments, etc.) due to weakened connections between representations. According to both accounts, partial access to phonological knowledge is accounted for by claiming partial activation of phonological representations, either as mediated by a lexeme node (WEAVE++) or
CHAPTER 2. PROPERTIES OF SPOKEN LANGUAGE PRODUCTION

content words that license their use (Ferreira, 2000; Levelt, 1989). The general idea is that a speaker might retrieve, say, a noun, which in turn will trigger the retrieval of knowledge (Ferreira, 2000) or execution of a procedure (termed indirect selection by Levelt, 1989) that retrieves the needed function words. Again, this amounts to a claim that a function word is not retrieved directly by meaning, but is instead mediated by the syntactic properties of content-word knowledge.

1.4. Assembling the Sounds of a Word

1.4.1. Property 10: The sounds of a word are assembled anew

A potentially counterintuitive idea is that the individual sounds of words are assembled anew each time they are spoken rather than retrieved as intact wholes. Yet, patterns of speech errors and latency data suggest that this is the case. According to one estimate, errors involving sounds occur approximately 2.6 times per 1000 sentences or 1.5 times per 10,000 words in spontaneous speech, whereas word errors occur at a rate of 4.4 per 1000 sentences or 2.5 per 10,000 words (Deese, 1984). When unimpaired speakers name isolated objects, errors involving the sounds of words are much less likely than word substitution errors (Dell et al., 1997). Sound errors include omissions, additions, and exchanges of individual sounds. The most common type of error is the anticipation of an upcoming sound (Nootseboom, 1973) as in alko share for also share (Fromkin, 1971).

Levelt and colleagues have argued that one reason that a word’s sounds must be assembled anew each time is due to changes in metric structure contingent on the accompanying words and inflections (e.g., Levelt et al., 1999). For example, whereas the /d/ is syllabified with the other sounds in hard when the word is spoken alone, it is syllabified with the following word, it, in the utterance Hand it [‘han-dit’]. The importance of metric structure can also be seen in benefits of repeating syllable structure independent of syllable content (Sevald, Dell, & Cole, 1995). That is, speakers can repeatedly produce kem-til-fer much faster than they can produce kem-til-lan, because the first two syllables of the first sequence share syllable structure (CVC) whereas the first two syllables of the second sequence do not (CVC and CVCC).

1.4.2. Property 11: Experience strongly affects speed and accuracy of assembling words

A striking fact about slips of the tongue is the way they reflect both long- and short-term experience with language patterns. Words fall apart in ways that reflect the sequences of sounds a speaker is most familiar with. Slips of the tongue are more likely to create words that exist in a speaker’s language rather than create novel sequences of sounds, a phenomenon known as lexical bias (Barts, Motley, & Mackay, 1975; Dell & Reich, 1981). Even when novel sequences are created, sounds in these new sequences only occur in syllable positions that they occupy in existing words of the language.

* There were controls for length and syllable complexity of course.
The sounds that slip tend to be those in the least predictable positions within the language of the speaker (Berg, 1998; Nooteboom, 1973). For example, word initial consonants (e.g., /b/ in the word bicycle: /baɪs/ekl/), are less predictable than other consonants that begin syllables in English (e.g., the /b/ and /k/ in bicycle), and in Germanic languages, slips of the tongue are more likely to separate a word initial consonant from a word than another syllable initial consonant (Schatruck-Hufnagel, 1987). This also shows up in games and sayings where everything except the initial consonant of a word is repeated (e.g., Pig Latin and "heiter skeleton" or "piggly wiggly"). Languages like Spanish that do not have this difference in distribution of word-initial vs syllable-initial consonants do not show this difference in phonological speech errors (Berg, 1991). Whole syllables often operate in speech errors in Mandarin Chinese (Chen, 2000, as cited by Chen, 2000, & Dell, 2002) but rarely in English. Likewise, in Semitic languages such as Arabic, where morphemes are discontinuous and the syllable positions of consonants change more often across words, speech error and poetic rhymes pattern very differently from languages like English where morphemes tend to be concatenated and maintain syllabic position across words (for comprehensive review, see Berg, 1998).

Even short-term experience with particular sound-ordering conventions affects the likelihood of making different types of errors. So, when in the context of a particular experiment, sounds only occupy particular positions, speakers' speech errors come to reflect these biases even when they are not part of the language in general. For example, when /l/ only occurs at the beginning of syllables in an experiment, speech errors involving /l/ nearly always involve the beginning of syllables (Dell, Rees, Adams, & Meyer, 2000). Even individual phonological features such as place of articulation (e.g., the difference between /b/, /d/; and /g/ are sensitive to these effects of experience (Goldrick, 2004).

Speakers appear sensitive to the frequency of whole words in addition to sequences of sounds within them. Unsurprisingly, children tend to learn common words earlier than uncommon words (e.g., Huttenlocher et al., 1991). Thus, it is difficult to determine whether it is the age at which a word is typically learned (its age-of-acquisition), how often it tends to be used (its word frequency), or both that affect word production (for a discussion of attempts, see e.g., Brysbaert & Ghyselinck, in press). More common or earlier learned words are generated as much as 100 ms rapidly than less common words (Oldfield & Wingfield, 1964). This speed advantage for common words may be due to the benefits of experience in word selection and phonological encoding, but several results suggest that the impact of frequency and age-of-acquisition is greater in phonological encoding than in word selection (see Brysbaert & Ghyselinck, in press). For example, lower word frequency increases the likelihood of phonological word substitutions, noise of the tongue (Dell, 1990; Steinberger & MacWhinney, 1986), and TOTs (Burke et al., 1991; Harley & Bown, 1998), but only seems to affect the likelihood of semantic word substitutions in unimpaired speakers when they are under heavy time pressure to speak (Vitkovitch & Humphreys, 1991).

Although they sound the same, the meanings and therefore the initial stages of word production differ for generating homophones, such as ball meaning a spherical object and ball meaning a formal dance. Given that they sound the same, one might suspect that their phonological processing, phonological representations, or both would be the same. Indeed, several experiments suggest that homophones share phonological processing and representations. Although ball meaning a formal dance is a very uncommon word, it is less prone to slips of the tongue than equally unusual words (Dell, 1990). In fact, providing anaphoric practice generating one meaning of a homophone carries over to improved performance in generating the other for a week (Biedermann, Blanken, & Nickels, 2002). Likewise, as described above, producing the meaning of one member of a homophone pair (pis as in cherry pit) makes a TOT less likely for the other member of the pair (e.g., Pis as in Brad Pitt; Burke et al., 2004). Hearing a word related to one homophone (e.g., dance related to the formal dance meaning of ball) speeds the naming of objects corresponding to the other homophone meaning (i.e., ball meaning a spherical object; Cutting & Ferreira, 1999) and speeds reading aloud the ambiguous printed word (e.g., ball; Balota & Paul, 1996). All of these results are readily accounted for by assuming that different meanings of homophones share a common representation of phonological form but differ in the lexical representations accessed through meaning (e.g., Dell, 1990; Jescheniak & Levelt, 1994). Alternatively, all lexical and morphological representations could be separate for the two homophone meanings, but bi-directional connections to shared sounds might lead to strong priming effects by allowing activation to spread back and forth between separate representations. Only Caramazza’s (1997) word production model explicitly excludes both shared representations for homophones and bi-directional activation, making it difficult to see how it could account for homophone effects. However, this feature of the model readily allows it to account for experiments suggesting that low-frequency homophones might not be produced as quickly as their high-frequency partners or their combined frequencies when important factors are controlled (Caramazza, Costa, Miozzo, & Bi, 2001; Jescheniak, Meyer, & Levelt, 2003).

Unlike the above-noted effects of the frequency of words or word patterns, the effect of the frequency of syllables upon production is less clear. Levelt and Wheeldon (1994) reported that the frequency of the final syllables of words influences production time independently of word frequency or whole-word naming time. More recent experiments showed that for disyllabic non-word production, Dutch-naming latencies (the same language assessed by Levelt and Wheeldon) were influenced by the frequencies of first syllables but not second syllables (Cholin, Levelt, & Schiller, in press). More work is necessary to sort out exactly when syllable frequency does and does not affect production times.
1.4.3. Property 12: Aspects of sound assembly proceed sequentially

Processing seems to start earlier in time for sounds at the starts of words than for sounds at the ends. For example, in picture-word interference tasks, distractors that share the initial sounds in object names have effects at earlier points in time than distractors that share later sounds of names (Meyer & Schrieffers, 1991; e.g., an initially overlapping word like tilt facilitates naming of a tiger at earlier points in time relative to a word like liar that only overlaps in final sounds). When speakers are told in advance that they will be asked to articulate one word out of a set that share initial sounds, they begin speaking earlier than if the word comes from a group that does not share initial sounds (Meyer, 1990, 1991). That is, speakers appear able to prepare the shared parts of the words in advance, leaving less material to prepare and allowing faster production than when nothing is known in advance about the form of the upcoming word. However, this foreknowledge is only helpful in Dutch when words share initial sounds and metrical structure (specifically, number of syllables and stress pattern; Roelofs & Meyer, 1998). In contrast, knowing in advance the final sounds or syllables of words provides no benefit (Meyer, 1990, 1991). Knowing just a phonological feature such as place of articulation at the start of a word also does not provide any detectable benefit, suggesting that the relevant level of preparation involved is whole phonemes or sounds (Roelofs, 1999). Somewhat similarly, generating a word with word-initial overlap (e.g., tile and tiger) slows naming of an object on a subsequent trial, whereas generating a word with word-final overlap (singer and tiger) speeds naming relative to generating an unrelated word (Wheelond, 2000). When speakers repeat word pairs multiple times, it takes more time per pair for combinations that share initial sounds relative to those that share no sounds, which in turn take more time than combinations that only share final sounds (Sevald & Dell, 1994). Although there are intriguingly different patterns of effects for hearing a word with overlapping initial segments vs. generating one, all of these results suggest a sequential process associated with retrieving, organizing, or programming speech sounds.

The WEAVER++ model covers all bases by having both simultaneous retrieval of all segments in a word, followed by a step in which each segment is associated with a syllable position in sequential order (Levett et al., 1999). Thus, speech processing in the model is sensitive to the availability of all phonemes in the first part of phonological encoding and the time needed to sequentially associate them with a syllable position. In addition, there is a final stage of phonetic encoding in which the phonologically specified syllables of words are sequentially associated with stored articulatory gestures. While having two stages that show sequential processing allows the model to account for sequential effects in production, it also makes it difficult for the model to simultaneously account for the absence of certain length effects. Specifically, object naming latencies and gaze durations on objects suggest that when the many potential confounds with word length are controlled, speakers take the same amount of time to prepare a multi-syllabic word as a monosyllabic one (Bachoud-Levi, Dupoux, Cohen, & Mehler, 1998; Bonin, Chalard, Meot, & Payol, 2002; Griffin, 2003; Sternberg, Knob, Monsell, & Wright, 1988; for discussion, see Meyer, Roelofs, & Levett, 2003).

1.4.4. Property 13: The effect of similar sounding words is highly situation-dependent

The effect of recent experience with a word that is phonologically similar to an intended word sometimes speeds (e.g., Starreveld, 2000) and sometimes slows word production (e.g., Wheelond, 2003), indicating that such effects depend on a complex set of factors. These differing effects may be due in part to experimental paradigms differentially calling on phonological subprocesses, such as sound retrieval as opposed to sound sequencing, and similar sounding words having different effects in sound retrieval, associating sounds with metrical structure, translating these phonological plans into motor programs, and articulation (see Levelt et al., 1999; O'Seaghdha & Marin, 2000). Complicating the interpretation of phonological priming effects in production (e.g., Starreveld, 2000), similar sounding words compete with one another in word recognition (e.g., Tanenhaus et al., 1995). In addition to the position in the words where sounds are shared (Sevald & Dell, 1994), the duration and type of processing that the first or priming word undergoes appears critical in shaping effects (O'Seaghdha & Marin, 2000).

When a speaker unintentionally asks for balaklava in a Mediterranean restaurant (rather than baklava), it is tempting to conclude that similar word forms compete against one another. Instead, speakers may say words that sound similar to their intended words as near-misses, in which they fail to retrieve all of the sound information for an intended word and default to a very similar form (Burke et al., 1991). In this category of effects, one can list the tendency for slips to be real words rather than novel sequences of sounds, the tendency for intruders in phonological word substitutions to have the same number of syllables and other characteristics as intended words (Fay & Cutler, 1977; Gagnon et al., 1997), the tendency for speakers in TOT states to often come up with similar sounding words (Burke et al., 1991), the tendency of slips of the tongue to involve sounds that share many phonological features, such as /h/ and /k/ rather than /h/ and /v/ (Franklin, 1971; Shattuck-Hufnagel & Klatt, 1979), and the tendency for sounds to exchange between similar sounding words (e.g., Dell & Reich, 1981). That is, words that sound alike do not appear to interfere and compete with one another during phonological encoding in the way semantically related words do in word selection. Indeed, although selecting a word from a semantically dense neighborhood seems to take more time than selecting one from a sparse neighborhood, the opposite seems to hold for phonologically defined neighborhoods. Words that share many sounds with other words take less time to generate than words that are more unusual (Vitevitch, 2002) and appear more likely to be successfully retrieved in terms of fewer phonologically related word substitutions (Vitevitch, 1997) and TOTs (Harley & Bow, 1998). Also supporting the idea that similar sounding words support each other rather than compete is the observation that priming with phonologically related words can resolve and prevent TOTs (James & Burke, 2000; Meyer & Bock, 1992). Likewise, presenting phonologically related distractor words during object naming speeds naming latencies relative to unrelated distractors (e.g., hill vs. ankle for a lion; Schriefes et al., 1990). Simulation studies conducted with interactive activation models suggest that feedback of activation from phonological neighbors may aid intended words in competing against their semantic neighbors (Dell & Gordon, 2003).
That said, there are situations in which having similar sounding words slows speech or increases the likelihood of errors. The most obvious case of such interference is in tongue twisters such as the sixth sick sheik's sixth sheep's sick. Experiments indicate that repeating words with similar initial sounds is significantly more difficult than repeating sequences with unrelated sounds (Sewald & Dell, 1994). In addition, speakers are slower to generate a name for an object (e.g., hod meaning hat in Dutch) when they generated a word with overlapping initial sounds (hond) on the preceding trial rather than an unrelated word (Wheldon, 2003). Dell and O'Seaghdha (1992) suggested that these and related phenomena reflect sequentially cued phonological competition, whereby having a sequence of phonemes (e.g., /h/) with one ending (/h/, in sheik), cues of the recently used ending makes it more difficult to subsequently complete that sequence (/h/) with a different ending (/p/, in sheep). Such sequential competition is readily accounted for with the subclass of connectionist models called simple recurrent networks and control signal networks that output phonological segments at a time for a given input (e.g., Dell et al., 1993; Vosden, Brown, & Harley, 2000). Although such models do an excellent job of producing some phenomena associated with phonological word assembly (particularly the effects of experience, similarity, and order on speech errors), it is unclear how they would be integrated with other parts of the production system to account for phenomena such as phonological influences on word selection.

1.5. Time course of processes in word production

1.5.1. Property 14: Semantic competitors activate their sounds

Despite the near consensus on the need for two stages to the production process, a famous controversy among theories of word production concerns the extent to which processing of sound and meaning overlap in time. In one manifestation of this, researchers have debated whether sound-related information is only processed after word selection is complete (e.g., Dell & O'Seaghdha, 1992; Dell & Reich, 1981; Harley, 1993; Levelt et al., 1991; Peterson & Savoy, 1998). On one side are models that characterize the flow of information during production as strictly staged—speakers first use activated meaning-level representations to perform word selection and only access sound information after the completion of the selection process. The most prominent model of this strictly discrete sort is the WEaver++ model presented in Levelt et al. (1999), which was developed computationally in Roelofs (1992, 1997). Other theorists have also argued for the strict separation of word selection and sound-processing stages (e.g., Butterworth, 1989; Caramazza, 1997). On the other side are models that assume staged processing, but allow activation to flow relatively freely among meaning, lexical, and sound representations, making multiple types of information relevant to both word selection and sound assembly (e.g., Dell, 1986; Harley, 1993). Specifically, partially activated but ultimately unselected lexical representations are permitted to influence sound assembly (via cascading activation). For example, before ultimately naming an object as couch, a speaker should activate both the word representation for couch and its synonym sofa (see Property 2, that speakers activate a family of meaning-related words) and, via cascading, the sounds of these words. Indeed, speakers are faster to read aloud words that are phonologically related to dispreferred synonyms of object names (e.g., soin for sofa when preparing to name a couch) when they are presented after beginning to prepare to name a drawing of a couch rather than a completely unrelated object (Peerson & Savoy, 1998; replaced by Jescheniak & Schriefers, 1997; see also Jescheniak & Schriefers, 1998). The WEaver++ model makes the post-hoc assumption that word selection is delayed until after sound processing begins only in the case of synonyms (Levelt et al., 1999).

Another manifestation of this controversy has concerned whether the sounds of ultimately unselected words may influence which word is selected. In models with bidirectional flow of activation or feedback, partially activated or ultimately unselected phonological representations are allowed to send activation backwards to affect lexical (and perhaps even semantic) levels of representation. The most prominent implemented model of this interactive sort is presented in Dell (1986), but this type of interactive activation has been incorporated in many theories and models (e.g., Dell et al., 1997; Eriksen, Schade, Kopiez, & Laubenstein, 1999; Harley, 1993; MacKay, 1982, 1987; Stemberger, 1983). Explaining the mixed error effect is one of the primary motivations for assuming this type of interaction. It turns out that the intruding words in semantically related word substitutions bear a greater than chance phonological similarity to the intended words that they replace (Berbaum & Valentine, 1992; Dell & Reich, 1981; Harley, 1994; Martin, Wexberg, & Saffran, 1989). In interactive-activation models with feedback, when generating the word cat, activation spreads to words related in meaning to cat such as dog, mouse, and rat, and via these word representations to their sounds. The sounds that form the word cat are highly activated by their link to cat's word node and they relay a portion of that activation to other words containing the same sounds such as cap, kit, and rat. Thus, a word that is both semantically and phonologically related to the intended word such as rat receives converging activation from both semantic and phonological representations, making it more likely to be selected by mistake than a word activated by only one of these sources. In contrast, discrete two-stage models account for mixed errors using an independently motivated error-checking mechanism (see e.g., Motley, Camden, & Baars, 1982). The basic idea is that the more a substituting word resembles an intended word, the less likely a pre-articulatory editing mechanism is to detect the error and prevent it from being uttered. Thus, under this account, mixed errors are not made disproportionately often, it is just that other errors are more likely to be detected and prevented, making the types of errors observed representative of those created in the language production system (Butterworth, 1982; Levelt, 1989). Through treatments of the issues of discreteness and interaction in word production can be found in Rapp and Goldrick (2000) and Vigliocco and Hartsuiker (2002).

1.5.2. Property 15: The scope of message planning is greater than the scope of sound assembly

Early in the study of speech error patterns, researchers noted that there was a greater distance between words that exchange places than between sounds that exchange places. For example, Nootbooom (1973) noted that 2 syllables separated exchanging sound
CHAPTER 2. PROPERTIES OF SPOKEN LANGUAGE PRODUCTION

before beginning a fluent utterance, the primary units proposed have been based on prosody or syntax. In the psycholinguistic literature, a phonological or prosodic word is typically defined as a single content word along with any adjacent, unstressed function words as in [beer's] [the good] [thing] (Ferreira, 1993; Wheeldon & Lahiri, 1997). Latency data indicate that speakers prepare at least one phonological word prior to initiating an utterance, with more complex phonological words delaying speech onset (Wheeldon & Lahiri, 1997). Other studies suggest that speakers will prepare more than one phonological word prior to speech if it does not form a whole lexical word (i.e., it is half of a compound; Wheeldon & Lahiri, 2002); if the second phonological word is part of the first noun phrase (Cava & Caramazza, 2002); and if the first word will not take long to articulate and speakers try to avoid pausing (Griffin, 2003). Strengthening the case for considering the phonological word an important unit at some level is the observation that the latency to begin articulating pre-planned speech is a function of the number of phonological words the pre-planned utterance contains (Snertberg et al., 1988; Wheeldon & Lahiri, 1997).

Other researchers have argued for phrase-wise word planning (e.g., Martin, Miller, & Yu, 2004). Certainly in many languages (e.g., Dutch, German, and Spanish), grammatical dependencies between nouns (e.g., beer, ale) and the adjectives (e.g., happy, amber) and determiners (e.g., a, some) that modify them make it necessary to retrieve information about the noun to determine the correct form of the adjective or determinant. Not surprisingly, picture-word interference studies suggest that nouns are selected before the onset of the determiner when speakers produce gender-marked determiner + adjective + noun phrases in languages such as Dutch and German (Schriefers, 1992; Schriefers, de Ruiter, & Steigerwald, 1999). Other work points to phrase-wise planning even in English speakers in the absence of strong grammatical dependencies. For example, English-speaking patients who, because of brain damage, have difficulty maintaining lexical-semantic information had greater difficulty producing utterances in which adjectives appeared in the same phrase as the noun they modified (e.g., the long, brown hair) than utterances in which the adjectives appeared in a different phrase (e.g., the hair was long and brown; Martin & Freedman, 2001). Tellingly, patients with impaired memory for phonological information did not show this difference and could produce these utterances as readily as unimpaired speakers.

There is mixed evidence for pre-speech planning of multiple nouns when they occur in a conjoined noun phrase such as monkey and chair. Support for phrase planning comes from finding of semantic interference effects on speaking latencies for both objects within a conjoined noun phrase (Meyer, 1996; but see Meyer, 1997) and when the nouns in the conjoined phrase name semantically related objects (Freedman, Martin, & Biegel, 2004). All else being equal, timing experiments indicate that speakers take about 70 ms longer to initiate sentences with two nouns in a conjoined subject noun phrase than sentences with a single noun (Martin et al., 2004; Smith & Wheeldon, 1999). Such observations have been used to argue that the contents of a noun phrase are processed in parallel (with a small cost) and that articulation of a sentence-initial conjoined noun phrase is not initiated until both nouns are prepared. In contrast, eye-tracking

---

1 A clause contains a verb and roughly corresponds to a simple sentence.
2 Shifting topics in a conversation is also associated with decreased fluency (e.g., Butterworth, 1975), suggesting that speakers engage in resource-demanding microplanning spanning multiple clauses.

segments, whereas 4.1 intervening syllables was the average distance between exchanging units of greater size such as morphemes and words. Such observations support the distinction between word and sound representations and separate processing stages that operate on them (e.g., Fromkin, 1971; Garrett, 1975). In addition, it suggests that abstract properties of words are specified further in advance than their sounds are.

Speakers seem to create proposition-sized pre-linguistic messages before they begin articulating an utterance. A rational argument is based on observing that the order of mention and syntactic structure of clauses depends on factors, such as the relative animacy, humanness, and agency of its message elements, suggesting that speakers may compare these prior to making any ordering or syntactic choices (e.g., Ferreira, 1994; Griffin & Bock, 2000). Empirically, speakers are more likely to pause, repeat words, and say um before articulating complex utterances and clauses than less complex ones (e.g., Beattie, 1979; Clark & Wasow, 1998; Ford, 1982). Syntactic complexity presupposes a complex message representation, so it is more parsimonious to attribute such effects to message planning than syntactic planning. Studies of simultaneous translators also suggest that speakers prefer to have a proposition-sized message prior to initiating an utterance (Goldman-Eisler, 1972). Speakers start long utterances with a higher pitch than shorter utterances (e.g., Cooper, Soares, & Reagan, 1985), indicating some degree of advance knowledge of content. At the start of a clause, speakers are slower to respond in secondary tasks (Ford & Holmes, 1978) and tend to avert their gaze from a listener (Kendon, 1967), also suggesting that they are engaged in more intense processing of some sort at these points in time. Likewise, equating for distance in words, nouns in the same clause are more likely to lead to subject–verb agreement errors than nouns in different clauses (Bock & Cutting, 1992), suggesting that items within the same proposition in a message are more available than those from different propositions.

Although speakers seem to know a lot about the message content of an utterance before they begin to speak, they often do not know all of the words they will use to express the message before they begin to articulate the utterance. Evidence from a wide variety of experimental tasks suggests that speakers often select the nouns that follow verbs while articulating the first words of their sentences (e.g., Griffin & Bock, 2000; Kempen & Huijbers, 1983; Lindsley, 1975; Smith & Wheeldon, 1999). However, variations in instructions or other aspects of a situation lead speakers to alter the number of words they prepare prior to speaking the first one (e.g., Griffin & Bock, 2000; Wheeldon & Lahiri, 1997). So, the question is what factors determine when speakers prepare and articulate their words?

There is a tradition in psycholinguistics of searching for warts in which planning is incremented. With respect to the minimum amount of planning a speaker must complete

---

1 A clause contains a verb and roughly corresponds to a simple sentence.
2 Shifting topics in a conversation is also associated with decreased fluency (e.g., Butterworth, 1975), suggesting that speakers engage in resource-demanding microplanning spanning multiple clauses.
experiments suggest that under similar circumstances, speakers prepare nouns one at a time, shortly before uttering them, even in complex subject noun phrases or conjunctions (Griffin, 2001; Meyer, Slezak, & Levelt, 1998; but see Morgan & Meyer, 2005).

An unanswered question is whether verbs (or other predicates) play a special role in the preparation of utterances. Based on the constraining properties of verbs, some theorists have suggested that verb selection must normally take place early in sentence formulation (e.g., Bock, 1987; Ferreira, 2000; Jarett, 1977; MacWhinney, 1987). When not required to select verbs in an utterance, speakers begin speaking earlier than they otherwise do (Kempen & Huijbers, 1983; Lindsley, 1975). Such results have been used to argue that verb selection precedes subject selection and therefore often speech onset (e.g., Bock & Levelt, 1994; Ferreira, 2000). However, these same experiments (Kempen & Huijbers, 1993; Lindsley, 1975) are also consistent with a desire to have a full or partially specified message planned before speech onset without verb selection, assuming that messages that include an action or other predicate take more time to compose, all else being equal, than those with only a topic. Similarly, the relationship between ear-to-mouth lag and verb position in translation input (Goldman-Eisler, 1972) supports the idea that a verb is selected before translated production begins, but also the more conservative possibility that production processes wait for a predicate to be included in the message. Further complicating matters is the possibility that speakers may only need to prepare verbs prior to speech onset whenever verbs occur soon after sentence onset (e.g., after short subject noun phrases in English) simply because there would not be time to prepare them while articulating the subject noun phrase (Griffin, 2003).

In addition to semantic and linguistic units and dependencies, time also appears to be important in timing speech. Longer words by definition take more time to articulate than shorter words, and slower speakers take more time to articulate their words than faster speakers do. Both of these aspects of timing have been shown to influence when speakers begin preparing words (Griffin, 2003). That is, speakers may attempt to minimize their buffering of prepared words by estimating how long words will take to prepare and how long it will take to articulate already prepared speech. Speakers are sensitive enough to the timing of articulation and word preparation that they will insert optional words such as that is The coach knew that you missed practice in response to variations in the availability of the following word (Ferreira & Dell, 2000). Also suggesting sensitivity to the time needed to prepare upcoming speech, speakers are more likely to say uh than um before shorter delays in speaking (Clark & Fox Tree, 2002).

2. SUMMARY

This chapter has described 15 basic properties of spoken language production. These properties characterize word production as consisting of a word-selection stage followed by a sound-processing stage (1). Selecting a content word such as a noun or verb involves activating (#2) and then competitively selecting (#5) from a family of meaning-related words in a grammatically constrained (#4) but meaning-sensitive (#5) fashion. This word-selection process may require an intention-to-name to have it commence (#6), and it may include an verb-generating component (#7). Nonetheless, word production can fail partly through (#8). Function words may undergo a somewhat different selection process than content words do (#9). Sound processing in turn is characterized as assembling sequences of sounds (#10), a process that is affected by speakers’ experience (#11), and proceeds from word start to end (#12). Phonological similarity has complex effects on production, attesting to the fact that it probably affects multiple subprocesses (#13). Although only one word may ultimately be spoken to produce a meaning, multiple meaning-related candidates can affect the availability of sound information (#14). Finally, speakers plan messages further in advance than they retrieve sounds, showing a tendency to prepare words for about a noun phrase at a time, due to message-level, syntactic, prosodic, and/or timing constraints or preferences (#15).

In focusing on spoken language and the production of words in particular, we have left untouched the literature on written language production (see e.g., Bork et al., 2002; Kellogg, 2003), the production of sign languages (e.g., Thompson, Emmorey, & Gollan, 2005), and the complexities of knowing words in multiple languages (e.g., Costa, Miozzo, & Caramazza, 1999; Gollan & Asenart, 2004; Kroll & Sunderman, 2003). Within spoken word production, this chapter has not addressed work on how speakers produce morphologically complex words (e.g., Badcock, 2001; Melinger, 2003; Roelefs, 1996; Wheelock & Lahiri, 2002), for discussion, see Waksler, 2000) such as morphology, laterbox, or ko-inutage (a Swedish word meaning "most tightly packed with cows") or idioms such as It's Greek to me and to put one's foot in one's mouth (see e.g., Cating & Bock, 1997; Levelt et al., 1999). We have barely touched on the production of prosody and the role of intonation in spoken language (for discussion, see Ferreira, 1993; Wheelock, 2000). Nor have we discussed under what circumstances and how speakers may or may not tailor their language to suit their audiences (Bar & Keesey, 2001; Ferreira & Dell, 2000; Ferreira, Slev, & Rogers, 2005; Horton & Gerrig, 2005; Kraljic & Brennan, 2005, Lockridge & Brennan, 2002). These are active and important areas of research in language production.

Most of the properties we have reviewed are sufficiently basic that they are virtually certain to characterize how production works, at least to some level of approximation. A few of them, however, are more controversial and are likely to be explained and revised by future research (e.g., whether the intention to name is critically involved in word activation [6], similarity in phonological encoding [12], and origins of phonological similarity effects in production [13]). Nonetheless, in all, these properties represent a tribute to the progress that the field of language production has made, as they represent true gains in our understanding of how speakers produce words. At the same time, they pose challenges to current and future models of production, as such models pursue their goal of transforming these descriptions of how production works into explanations of how it works the way it does.
REFERENCES


CHAPTER 2. PROPERTIES OF SPOKEN LANGUAGE PRODUCTION


Griffin, Z. M. (2004). The eyes are right when the mouth is wrong. Psychological Science, 15, 814–821.


CHAPTER 2. PROPERTIES OF SPOKEN LANGUAGE PRODUCTION


CHAPTER 2. PROPERTIES OF SPOKEN LANGUAGE PRODUCTION


