Speech Production, Psychology of

Herbert Schriefers, Nijmegen University, Nijmegen, The Netherlands
Gabriella Vigliocco, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

This article is reproduced from the previous edition, volume 22, pp. 14879–14882, © 2001, Elsevier Ltd.

Abstract

Research on speech production investigates the cognitive processes involved in transforming thoughts into speech. This article starts with a discussion of the methodological issues inherent to research in speech production that illustrates how empirical approaches to speech production must differ from related fields like language comprehension. Then, an overview of the processing stages engaged in production as generally agreed upon by the different theories in the field is presented. The processing stages are: conceptual preparation – turning a thought into a verbalizable message to be expressed; grammatical encoding – developing a syntactic frame for the to-be-uttered sentence; and phonological encoding – realizing the phonological content of the syntactic frame. For each processing stage, basic theoretical assumptions, representative findings, and open issues are discussed.

Speech production refers to the cognitive processes engaged in going from mind to mouth (Bock, 1995), that is, the processes transforming a nonlinguistic conceptual structure representing a communicative intention into a linguistically well-formed utterance. Within cognitive psychology, research concerning speech production has taken various forms such as: research concerning the communicative aspect of speaking; research concerning the phonetics of the produced speech; and research concerning the details of the cognitive processing machinery that translates conceptual structures into well-formed linguistic utterances. We focus on the latter.

Native adult speakers produce on average two to three words per second. These words are retrieved from a lexicon of approximately 30 000 (productively used) words. This is no small feat: producing connected speech not only entails retrieving words from memory, but further entails combining this information into well-formed sentences. Considering the complexity of all the encoding processes involved, it is impressive that we produce speech at such a fast rate while at the same time remaining highly accurate in our production. Bock (1991) estimated that slips of the tongue occur in speech approximately every 1000 words despite the ample opportunities for errors. In the following sections the cognitive processes subserving this ability will be discussed.

Methodological Issues

The prototypical empirical approach in cognitive psychology consists of systematically varying some properties of a stimulus (input), and to measure a corresponding behavior. Systematic relations between input and behavior are then used to infer the cognitive processes mediating between them. For example, researchers interested in language comprehension may systematically vary properties of a linguistic input such as the syntactic complexity of sentences or the frequency of words, and measure a corresponding behavior, such as reading times. Differences in the latter are taken as an indication of differences in the processes engaged in building an interpretation of a sentence (the output). Thus, the input is directly observable and completely accessible to systematic manipulation, while the output and the cognitive processes leading to it are inferred from the behavior.

The situation is different for research in speech production. The input (the conceptual structure to be expressed), is not directly observable and not readily accessible for experimental manipulation. By contrast, the output (i.e., the spoken utterance), is directly observable. Given this situation, research on speech production has started by focusing on properties of the behavior. Most prominent in this respect are analysis of speech errors, and of hesitations and pauses as they occur in spontaneous speech. These analyses motivated the general theoretical framework for speech production (Figure 1) which is described in detail below. More recently researchers have started to make speech production accessible to standard experimental approaches from cognitive psychology (see Bock, 1996).

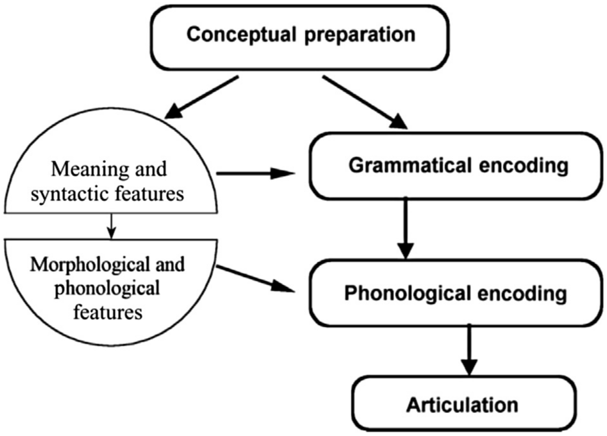

Figure 1 Levels of processing in speech production. Left part refers to stored representations, right part refers to processes integrating these representations into well-formed linguistic utterances.

Levels and Processes

Figure 1 provides an outline of the levels of processing involved in speech production. These levels are shared by most current psycholinguistic models (e.g., Dell, 1986; Garrett, 1988; Levelt, 1989). *Conceptual preparation* refers to the processes that transform thoughts into verbalizable messages. *Grammatical encoding* refers to the processes that turn the messages into syntactically well-formed skeletons for sentences. *Phonological encoding* refers to the processes that provide flesh to the syntactic skeletons, realizing the phonological content of the sentences. At these latter two processing levels, stored lexical information is retrieved from a long-term memory store, the mental lexicon. In particular, during grammatical encoding it is assumed that abstract lexical representations, specifying the syntactic properties of the words, but not their phonological content, are retrieved. During phonological encoding, word form information is retrieved. Finally, *articulation* refers to the processes that implement the phonetic content and turn it into motor commands for speech.

These processing levels were originally inferred from properties of speech errors (*see* Speech Errors, Psychology of). Speech errors can be analyzed with respect to the type of linguistic unit involved (words, morphemes, phonemes), and with respect to constraints that hold for errors concerning a given type of unit. For example, word exchanges (e.g., 'on the room to my door' instead of 'on the door to my room') are constrained by syntactic factors (like the syntactic category of the exchanged words) whereas phoneme exchanges (e.g., 'heft lemisphere' instead of 'left hemisphere') are constrained by phonological factors (like their position as onset, nucleus, or coda of a syllable). These and related results suggest a distinction between a processing level that is primarily concerned with syntactic processing (grammatical encoding), and another level primarily concerned with phonological processing (phonological encoding). The fact that phonemes are relevant units in speech errors also shows that the phonological form of a word is not stored and retrieved as a whole, but built from smaller units, namely the phonemes.

Planning of a sentence is not completed on one processing level and then sent to the subsequent level. For example, phonological encoding starts as soon as the syntactic structure of some part of a sentence has been determined, while the grammatical encoding of the remainder of the sentence continues. This piece-by-piece generation of a sentence is referred to as incremental production. Although there is no consensus on the size of the relevant planning units at the different processing levels, it is generally agreed upon that the size of the planning units becomes smaller as one moves from conceptual preparation through grammatical encoding to phonological encoding and articulation. Incrementality provides a natural account of fluency in speech production. Furthermore, incremental production reduces the burden of storage of intermediate results of the production process at each level.

Although there is general agreement on the levels of processing, theories differ with respect to how the flow of information between levels is characterized. According to modular views, each level of processing first produces a complete output representation for a given planning unit. Only then is this output representation passed onto the next level (e.g., Levelt, 1989). According to interactive views, every processing stage can pass partial results to the next processing level, before the output representation at the higher level has been completely specified, and lower levels of processing can provide feedback to higher levels of processing. As a consequence, processes at the phonological level can affect processes at the grammatical level (e.g., Dell, 1986). In the following sections, we discuss the different processing steps in more detail.

Conceptual Preparation

A first important goal of conceptual preparation is to establish which parts of the conceptually available information are going to be encoded, and in what order (see Levelt, 1989, for an overview). A second goal is to convert the conceptual information into a format that is a suitable for the linguistic formulation processes. One important open issue concerning conceptual preparation is how to characterize the mapping between conceptual and grammatical encoding. First is the question of whether conceptual preparation takes language specific properties into account (*see* Language and Thought: The Neo-Whorfian Hypothesis). Languages differ in which conceptual or formal properties need to be realized as a detail of the sentential form. For example, in English the word 'friend' does not carry information concerning the sex of the friend. In Spanish the corresponding word is differentially inflected for a man ('*amigo*') or a woman ('*amiga*'). In English, adjectives used as predicates (e.g., 'tall' in 'The friend of Luis is tall') do not agree in gender with the noun; in Spanish they do (e.g., '*El amigo de Luis es alto*' or '*La amiga de Luis es alta*'). Thus, in these two languages, conceptual information concerning natural gender must (Spanish) or need not (English) be conveyed by a sentence. Second is the question of whether conceptual information permeates the processes occurring during grammatical encoding beyond providing its input (Vigliocco and Franck, 1999).

Grammatical Encoding

Grammatical encoding refers to the processes involved in developing a syntactically well-formed sentence. It comprises first, those processes that map the relationships among the participants in a conceptual representation (e.g., agent, patient, etc.) onto functional syntactic relations between the words of a sentence (e.g., subject, direct object, etc.). Next, on the basis of the resulting hierarchically organized syntactic frame for the sentence, the words in the sentence are linearized in a manner allowed by the language being spoken. The first step is also referred to as *functional level processing*, the second as *positional level processing* (Garrett, 1988). Distinguishing between building hierarchical and linear frames provides a solution to an important problem that the language production system faces. Speech production has to be incremental to allow for fluent utterances, but at the same time the resulting utterance has to obey language specific constraints that force the use of only certain word orders. Assuming incremental conceptualization, however, the order in which parts of a conceptual message are processed do not necessarily correspond to a word order allowed in the speaker's language. This problem can be solved by separating the construction of hierarchical structures from the serial ordering of the words. In this way, hierarchical structures can be built in an incremental manner as soon as lexical elements are available, these can then be mapped to permissible linearly ordered positions (Vigliocco and Nicol, 1998). Evidence compatible with such a separation comes from studies showing that the linear position of words can be primed by the previous presentation of the same linear order, even if the hierarchical structure differs (Hartsuiker et al., 1999).

An important open question with respect to grammatical encoding concerns whether the encoding at this level can be influenced by feedback from phonological encoding (Bock, 1986). Research on the process of computing number agreement between a sentence subject and a verb suggests that this grammatical process is sensitive to whether number information in the subject noun phrase is overtly realized in the phonological form of the noun, suggesting the possibility of feedback from phonological encoding to grammatical encoding (e.g., Vigliocco et al., 1995; but see Section Self-monitoring and Repair for an alternative interpretation of the results).

Phonological Encoding

Phonological encoding refers to the processes that are responsible for determining the phonological word forms and prosodic content of the sentence (e.g., Levelt, 1999). First, the phonemes of words are retrieved from the mental lexicon, together with a metrical frame which specifies stress pattern and number of syllables in the word. Following this retrieval process, the resulting sequence of phonemes is syllabified according to a language specific set of syllabification rules. The domain of syllabification is assumed to be the phonological word which can but need not coincide with a lexical word. It should be noted that in this view, the syllabic structure of an utterance is computed on-line. The reason for this assumption is that the actual syllabification of a word in running speech depends on the context in which it appears. For example, the word 'deceive' in isolation is syllabified as 'de-ceive,' but in the context of the utterance 'deceive us,' the syllable structure becomes 'de-cei-veus.' This observation also provides a functional reason why the phonological form of a word is not stored and retrieved as one entity: such a representation would have to be broken up into its constituent parts whenever the stored syllabification of a word does not agree with its syllabification in the context of running speech. The representation formed by phonological encoding, a syllabified phonological code, forms the input for the articulatory processes which realize this code as overt speech. It is an open issue whether the transition from phonological encoding to articulation involves accessing a 'syllabary,' i.e., memory representations specifying the motor tasks that have to be performed to generate each syllable (Levelt, 1999).

Lexical Access

Grammatical and phonological encoding make use of stored representations of words. Parallel to the distinction between grammatical encoding and phonological encoding, two levels of stored representation are distinguished. On one level, words are represented as abstract syntactic (and semantic) entities (technically also referred to as lemmas), and on the other level their phonological form is specified. Evidence for two levels of representation in the mental lexicon comes from tip-of-the-tongue (TOT) states. It has been demonstrated repeatedly that speakers in a TOT state can report grammatical properties of a word (e.g., the grammatical gender of a noun) above chance level even when they are unable to access the phonological form of the word.

It is unresolved whether, within the lexicon, the flow of information from meaning to form can be characterized as discrete, cascading, or fully interactive. Discrete serial models of lexical access (e.g., Levelt, 1999) assume that, first, on the basis of some fragment of a conceptual structure, a set of lemmas becomes activated. This set includes the target lemma as well as semantically related lemmas. Phonological encoding of the word is initiated only after the target lemma has been selected. Hence, the phonological forms of words which are semantically related to the target word should not become activated. Cascading models (e.g., Peterson and Savoy, 1998) assume that every activated lemma sends some activation to its corresponding word form, even before the actual target lemma has been selected. Thus, also the phonological forms of semantic competitors of the target word should receive some activation. Finally, interactive models (e.g., Dell, 1986) allow for feedback from the phonological level to the lemma level. Thus not only will the phonological forms of semantic competitors receive activation, but the phonological forms will also send activation back to all lemmas with similar phonological forms.

The question of whether the phonological forms of semantic competitors do receive activation, as predicted by cascading models and feedback models, or not, as predicted by discrete serial models, is still an empirically unresolved issue. Phonological activation of semantic competitors has been demonstrated for the case of near-synonyms like 'couch' and 'sofa' (e.g., Jescheniak and Schriefers, 1998; Peterson and Savoy, 1998) and has been interpreted in favor of cascading of activation. However, as the evidence is presently restricted to this specific semantic relation, it has been suggested that near-synonyms might be a special case (e.g., Levelt, 1999).

Concerning the question whether there is feedback from lower levels to higher levels in lexical processing, the lexical bias effect has been cited as evidence for feedback. The lexical bias effect concerns the observation that phoneme errors lead to existing words rather than nonwords more often than would be expected by chance. This finding can be explained by feedback from the level of phonological segments to some higher level representing words as one integral unit (e.g., Dell, 1986). According to discrete serial models, by contrast, the chance of incorrect selection of a phoneme should be independent of whether the resulting string of phonemes forms an existing word or not (but see Section Self-monitoring and Repair).

Self-Monitoring and Repair

Speakers not only produce speech, but also monitor their own speech output and this monitoring may occur on overt speech as well as on 'internal speech.' Assuming an internal monitor implies that not all misfunctions of the speech production system will become visible in the eventual output. Rather, some will be discovered and repaired before they are actually articulated. Such covert repairs can play an important role in the interpretation of speech error results. For example, the lexical bias effect in phoneme errors has also been explained as the result of a pre-articulatory covert repair mechanism. This covert repair mechanism might be more likely to detect and correct phoneme errors leading to nonwords than those leading to words before the resulting error is actually produced. Hence, the lexical bias effect would not imply feedback of activation from lower to higher levels. It should be noted that related arguments have been put forward with respect to other empirical findings which appear to speak against discrete serial models, like the potential influence of phonology on grammatical encoding processes. It is a matter for future investigation to design empirical tests that allow us to differentiate between feedback and monitoring explanations of the relevant empirical phenomena.

See also: Aphasia; Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Speech Errors, Psychology of; Speech Production, Neural Basis of.

Bibliography

Bock, J.K., 1986. Meaning, sound and syntax: lexical priming in sentence production. Journal of Experimental Psychology: Learning, Memory and Cognition 12, 575–586.

Bock, J.K., 1991. A sketchbook of production problems. Journal of Psycholinguistic Research 20, 141–160.

Bock, J.K., 1995. Sentence production. From mind to mouth. In: Miller, J.L., Eimas, P.D. (Eds.), Handbook of Perception and Cognition, Speech, Language, and Communication, second ed., vol. 11. Academic Press, San Diego, CA, pp. 181–216.

Bock, J.K., 1996. Language production: Methods and methodologies. Psychonomic Bulletin and Review 3, 395–421.

Dell, G.S., 1986. A spreading-activation model of retrieval in sentence production. Psychological Review 91, 283–321.

Garrett, M.F., 1988. Processes in language production. In: Nieumeyer, F.J. (Ed.), Linguistics: The Cambridge Survey, Biological and Psychological Aspects of Language, vol. 3. Cambridge University Press, Cambridge, UK, pp. 69–96.

Hartsuiker, R.J., Kolk, H.H.J., Huiskamp, P., 1999. Priming word order in sentence production. Quarterly Journal of Experimental Psychology 52A, 129–147.

Jescheniak, J.D., Schriefers, H., 1998. Serial discrete versus cascaded processing in lexical access in speech production: Further evidence from the co-activation of near-synonyms. Journal of Experimental Psychology: Learning, Memory, and Cognition 24, 1256–1274.

Levelt, W.J.M., 1989. Speaking. From intention to articulation. MIT Press, Cambridge, MA.

Levelt, W.J.M., 1999. Models of word production. Trends in Cognitive Science 3, 223–232.

Peterson, R.R., Savoy, P., 1998. Lexical selection and phonological encoding during language production: Evidence for cascaded processing. Journal of Experimental Psychology: Language, Memory, and Cognition 24, 539–557.

Vigliocco, G., Franck, J., 1999. When sex and syntax go hand in hand: Gender agreement in language production. Journal of Memory and Language 40, 455–478.

Vigliocco, G., Nicol, J., 1998. Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? Cognition 68, B13–B29.

Vigliocco, G., Butterworth, B., Semenza, C., 1995. Constructing subject–verb agreement in speech: The role of semantic and morphological factors. Journal of Memory and Language 34, 186–215.