

Where do spoken words come from?

Core operations in normal speech production are the accessing of words in memory that appropriately express the intended message, and the preparation of each word retrieved for articulation. The theory developed in the MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS, Nijmegen, The Netherlands, provides a detailed account of both mechanisms (PNAS, 98, 23, November 06, 2001).

Every normal person learns to speak, and speaking involves, among other things, producing words. By reaching adulthood a speaker in our Western culture may well have produced some 50 million words. There is hardly any other human skill that is so well practiced. In normal speech we produce words at rates of some 2 to 4 per second. These words are continuously selected from a mental lexicon containing tens of thousands of words. Still, we make few errors. On average, we select the wrong word (for instance left when we mean right) no more than once in a thousand items. How is this robust, high-speed mechanism organized?

The theory proposed consists of two major processing components (Fig. 1). The first component deals with lexical selection. It is the mechanism that, given semantic input (some state of affairs to be expressed), selects one appropriate lexical item from the mental lexicon. The second component deals with form encoding. It computes the articulatory gestures needed for the articulation of the selected item. The theory has been computationally implemented under the name of WEAVER ++, and its experimental verification involved a decade of teamwork by Levelt's research unit at the Max Planck Institute, particular involving Drs. Antje Meyer and Ardi Roelofs.

A major experimental paradigm used has been picture naming. A picture to be named, for instance one of a horse, is presented to a subject. The instruction is to name the picture as fast as possible. We measure the latency from picture onset to the onset of articulation. This latency is about 600 milliseconds for naming a horse.

Apparently, lexical selection and form encoding can be completed within two-thirds of a second. During lexical selection two successive operations are run. The first one, perspective taking, consists of selecting the target concept for expression. Experimental conditions can be manipulated such that subjects will use horse or animal or stallion to refer to the object. The second one, lemma selection, consists of selecting the one corresponding item from the mental lexicon, for instance the item 'horse'. The item is called a 'lemma', which roughly means 'syntactic word', i.e. the word's syntactic properties, such as word class (noun, verb, etc.) or other syntactic features (such as gender for nouns, or transitivity for verbs). Selecting the item appropriate to the target concept takes place under competition. Semantically related items, such as 'animal' or 'stallion', are measurably coactivated. The quantitative computational theory predicts the selection latencies for selection under competition. The theory is tested by way of picture naming experiments where subjects are presented with an auditory or visual distracter word while they name the picture. The dis-

tracter is to be ignored. If horse is the target name, hearing the unrelated word chair slows down the response latency by some quantity. But hearing the semantically related word 'cow' has an even stronger inhibiting effect, an extra 50-100 milliseconds (dependent on conditions). This so-called semantic inhibition effect has been tested and quantitatively confirmed in a large variety of experiments. The time course of lemma selection, predicted by the theory, was further tested and confirmed by way of magnetic encephalography (MEG) in joint work with the Max Planck Institute for Cognitive Neuroscience in Leipzig. That work showed, in addition, the involvement of regions in the left lateral temporal lobe in the operation of lemma selection. Form encoding is initiated upon selection of the target lemma. The first step here is the retrieval of the target item's phonological code, an abstract string of phonological segments, for instance (h, o, r, s). Retrieving a word's phonological code is faster for words that are frequently used than for low-frequency words (by some 40 milliseconds). In picture naming, retrieving the code can be facilitated by providing the subject with a phonologically related distracter word. Subjects are faster in naming a horse if presented with a distracter such as 'horn' than with one such as 'chair' (phonologically unrelated to target). The time course of phonological facilitation is exactly predicted by Roelofs's WEAVER++ model.

Upon retrieval of the code from the mental lexicon, the next operation is initiated, syllabification. Segments are incrementally concatenated to form syllables. Concatenating the segments h, o, r, s produces the phonological syllable /h o r s/. But if the target word would have been the plural noun (for instance when there were two horses on the picture), a disyllabic syllabification would have resulted: /h o r / - / s w z/. Hence, syllabification is context dependent. Whether the syllable /h o r s/ or /h o r / will be generated depends on the following context. Segmental concatenation in syllabification runs at a rate of about 25 milliseconds per segment. Incremental syllabification predicts a word length effect, confirmed by recent experiments: naming latencies are shorter for monosyllabic than for disyllabic words.

The final step in form encoding is phonetic encoding, the retrieval of articulatory scores for each of the incrementally generated syllables. The theory assumes the existence of a mental syllabary, a repository of syllabic gestures, motor programs for frequently used syllables. There is evidence for the involvement of premotor cortex/Broca's area in the storage of these over-learned syllabic gestures. The factual execution of these gestures by the laryngeal and supra-laryngeal articulatory system generates the overtly spoken word. But that is beyond the present theory.

