

MULTIPLE LEXICAL ACCESS IN SPEECH PRODUCTION

Willem J.M. Levelt

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT

It is quite normal for a fluent speaker to produce two to four words per second. The accessing system is robust; lexical error rates are far below 1%. How is such fast, repeated lexical access organized? I will first present a bird eye's view of the architecture of single-word lexical access, as emerging from speech latency studies, such as picture/word interference and implicit priming experiments. Preparing a content word involves a system consisting of two major components: the first one, *lexical selection*, is conceptually driven and produces a syntactically specified lexical item as output. The second component, *form encoding*, retrieves the item's phonological code, incrementally generates the item's syllabification in context and computes the corresponding articulatory gestures. These gestures are executed during articulation.

Starting from this architecture, I will begin discussing the process of multiple access from the analysis of simple two-content word utterances such as *scooter and camel*. I will contrast three possible planning structures: level-by-level, incremental, and unit-by-unit. Level-by-level: first do lexical selection for the whole utterance, then form encoding for the whole utterance, then initiate articulation. Incremental: do form encoding of word-1 while selecting word-2, or articulate word-1 while doing the form encoding of word-2, etc. Unit-by-unit: First go through all encoding steps for word-1, then through all encoding steps for word-2 (etc.). These procedures make different predictions with respect to the fluency of the resulting utterance, and with respect to the effect of word-2 related distracter words on speech latency. The level-by-level planning procedure, suggested in the early days of modern psycholinguistics, is quite untenable and will not be further discussed. Although most data support the incremental structure of utterance planning (introduced a quarter century ago by Dennis Fry), a small adaptation of the unit-by-unit planning structure generates exactly the same latency and fluency predictions. A different type of experimental data is necessary to distinguish these latter two procedures. Eye tracking registration during description of multiple-object pictures can do the job. The tracking data support the unit-by-unit rather than the incremental type of planning. Further theoretical analysis shows that unit-by-unit planning, different from incremental planning, circumvents buffering. However, not in all cases. When the articulation of word-1 is relatively short as compared to the preparation of word-2, articulatory buffering cannot be evaded if within-utterance fluency is to be preserved. Recent counterintuitive data by Griffin support this analysis. In conclusion, multiple word planning strategies are shaped by fluency requirements, by a general aversion against (articulatory) buffering, and by the relative durations of articulation and preparation of successive words.