Phonological encoding in speech production: Comments on Jurafsky et al., Schiller et al., and Heuven and Haan

Willem J. M. Levelt

1. Introduction

What is phonological encoding? An introductory answer to this question may be helpful for a discussion of the papers in the present section. The term 'phonological encoding' has multiple uses, as Pat Keating signaled during the meeting from which this book stems in her introductory presentation: it can denote the encoding *by* phonology or the encoding *of* phonology. In the reading literature, for instance, the standard use of the term is this: 'Phonological encoding is writing a letter or word based on its sounds'[1] i.e., the encoding of phonology *by* orthographic units. This is not the topic of the present section. The other use of the term, encoding *of* phonology was introduced in my book *Speaking* (1989). After a discussion of grammatical encoding, phonological encoding was introduced as follows:

"Second, there is *phonological encoding.* Its function is to retrieve or build a phonetic or articulatory plan for each lemma and for the utterance as a whole. The major source of information to be accessed by the Phonological Encoder is *lexical form,* the lexicon's information about an item's internal composition. Apart from the lemma information, an item in the lexicon contains information about its morphology and its phonology [...] Several phonological procedures will modify, or further specify, the form information that is retrieved. [...] The result of phonological encoding is a *phonetic* or *articulatory plan.* It is not yet overt speech; it is an internal representation of how the planned utterance should be articulated - a program for articulation." (p. 12)

Introducing this use of the term to denote the speaker's phonological preparation of the utterance was by no means introducing the study of this process. Clearly, the study of phonological encoding has its roots in speech error analysis, going back as far as Meringer and Mayer's (1895) careful analysis of form errors in spontaneous speech. They distinguished between meaning-based substitutions (such as *Ihre* for *meine)* and form-based substitutions (such as *Studien* for *Stunden),* which suggested the existence of two levels of processing in speech production, one of which concerns form encoding. This notion was elaborated in much detail by Garrett (1975). He discovered that word exchanges (such as *he left it and forgot it behind)* can be between phrases or clauses, but preserves grammatical category and grammatical function. By contrast, sound exchanges (such as *rack pat* for *pack rat)* mostly happen between juxtaposed or close-by words, which can differ in grammatical category and function. Apparently two levels of processing are involved in utterance generation, which Garrett called the 'functional' and 'positional' levels. The latter involves morphological processes such as affixation and all further phonological (though not phonetic) processing. Except for details of phonetic preparation this level of processing coincides with 'phonological encoding' as outlined above.

A landmark development in modeling the process of phonological encoding was Shattuck-Hufnagel's (1979) Scan-Copier Model, with its slots-and-filler mechanism of word form encoding in utterance context. The review of phonological encoding in Levelt (1989) is largely inspired by this modeling effort. The Scan-Copier Model was based on both a detailed corpus analysis of phonological speech errors and on the results of systematic error-inducing experiments, an important methodological innovation, which brings us to laboratory phonology.

In retrospect, the laboratory study of phonological encoding was quite late to develop. Whereas the experimental study of speech perception was a long established field by 1975, the laboratory study of speech production was either articulatory phonetics, the study of the vocal tract's production of speech sounds, or reading-based study of prosody. There existed a tacit but quite general

disbelief that one would ever be able to gain experimental output control over a speaker's natural utterance production, including phonological encoding.

This has drastically changed. A range of experimental paradigms have been invented over the past quarter century that do provide for that type of output control. Among them are the above-mentioned error-induction paradigms, introduced by Baars et al. (1975). But more importantly, there are the chronometric naming paradigms stemming from an old tradition in studies of reading and picture naming (cf. Levelt, 1999 for a review). That tradition first touched issues of phonological encoding when Lupker (1982) discovered that the latency of naming a picture was reduced when simultaneously with the picture a visual distractor word was presented that rhymed with the target picture name (as compared to a situation where a non-related distractor word was presented). This 'orthographic' facilitation is really phonological facilitation. It is also obtained when the distractor word is presented auditorily. Any segmental phonological correspondence between distractor and target can induce facilitation (Schriefers et al., 1990; Meyer & Schriefers, 1991). The contribution of Schiller et al. in the present section uses orthographic picture/word interference to study, cross-linguistically, whether syllable priming exists in phonological encoding.

Together with picture/word interference, a range of other chronometric paradigms were developed to study phonological encoding (see Levelt 1999 for review), among them the 'implicit priming paradigm' (Meyer, 1990). If subjects are induced to produce a block of words that are word-initially phonologically identical (such as *loner, local, lotus* - 'homogeneous' block), word onset latencies are shorter than when the same words are produced in 'heterogeneous' blocks (e.g. the word *loner* among *beacon* and *major).* The shared word onset in a homogeneous block is an 'implicit' phonological prime. The method has been profitably used to study various issues of phonological encoding, among them: Is phonological encoding an incremental procedure? The clear answer is 'yes'. Only word initial primes are effective. Sharing word final phonology (as in *deed, feed, seed)* is entirely ineffective. This

doesn't contradict the just mentioned rhyme priming effects in picture/word interference. These arise at an earlier level of the encoding process. In phonological encoding a first step is to 'spell out' the target word's phonological code, largely its segmental composition; it is most unlikely that the stored code has syllable structure, because syllable structure can be highly context dependent (Levelt, 1992). Typically a word's segments are simultaneously activated; the spell-out can be primed through any segment in the code. During a next stage the code is used to compute the phonological word proper. This is an incremental process of syllabification. The domain of syllabification can be smaller (compounds) or larger (clitics) than the input lexical word. And syllabification is incremental. It starts word-initially, synthesizing one syllable after another. This process is the one accessed by the implicit priming paradigm. The paradigm has also been used to show that the elements of incremental encoding are segments, not features. Implicit priming is back to zero when the word initial segments in a block differ by just one feature, as in *noga, modem* (Roelofs, 1999). In a modified form the paradigm has further been used to study metrical aspects of word encoding. The evidence (reviewed in Levelt et al. 1999) supports the notion that, at least for Dutch, a phonological word's stress pattern is computed 'on-line'. It is probably the case that only irregularly stressed words carry metrical information in their stored phonological codes.

The two ordered operations of retrieving the stored phonological codes and using them in incremental prosodification (syllabification, metrical encoding, as well as higher levels of prosodification) are followed by phonetic encoding. As incremental prosodification proceeds, the resulting syllabic and larger prosodic structures should acquire phonetic shape. How do speakers incrementally prepare the articulatory gestures for the subsequent syllables in their prosodic contexts? One hypothesis, originally proposed by Crompton (1982) and further developed in Levelt (1989) and Levelt and Wheeldon (1994), is that frequently used syllabic gestures are stored as abstract motor schemas, somewhere in premotor/Broca cortex. This hypothetical repository has been called the 'mental syllabary'. And indeed, as Schiller has shown (see Levelt

et al. 1999), English and Dutch speakers do more than 80% of their talking with no more than 500 different syllables. My rough estimation is that, on average, we have used each of these syllables almost 100.000 times at reaching adulthood, i.e., some 13 times every single day[2].

Many of these syllables are themselves high-frequency words and there is no reason why even multisyllabic high-frequency words (such as *about* or *really)* wouldn't be similarly stored in this speech motor repository. Levelt et al. (1999) suggest a mechanism by which these high-frequency target syllables are incrementally selected, as phonological syllabification proceeds. This cannot be the full story, of course. Speakers are also able to phonetically encode low-frequency and even new syllables. In addition, phonetic encoding involves the further coarticulatory integration of successive articulatory syllables.

The three papers discussed here address different aspects of phonological encoding. Jurafsky et al. focus on the issue of modularity in phonological encoding: do homophonous words behave similarly in phonetic encoding, as the above theory predicts, or is their phonetics co-determined by their specific lexical frequency? As mentioned, Schiller et al. consider whether accessing the phonological code already involves accessing syllable structure. Van Heuven et al., finally, address higher levels of phonological encoding and decoding, relating to intonational accentuation in statements versus questions. I will now consider these contributions in turn.

2. Commentary

The empirical basis of the staged model of phonological encoding, summarized above, is formed by chronometric laboratory data, mostly response latencies in word and phrase production experiments. In their article, Jurafsky et al. managed to test aspects of that theory against the wider empirical domain of naturalistic speech data. If homophones, such as the pronoun *that* and the complementizer *that,* share their phonological code but not their lemma, as Levelt et al. (1999) claim to be the case, could one still

observes articulatory differences between them in natural connected speech? Strictly speaking, the theory and in particular Roelofs's (1997) computational model WEAVER++, predict only that there will be no difference in latency. Latency differences, however, are exactly what is hard to observe in naturalistic data; one has no natural anchor point for word onset latency. Jurafsky et al. checked instead how the word is realized, its duration, its vowel quality, its coda. It conforms to the nature of the theory to predict that homophones are phonetically realized in the same way (but see below for a qualification). In particular, the frequency of the lemma should be irrelevant. If not, one has a so-called 'lemma effect'. Of course, one should correct for confounding factors, such as position of the word in the utterance, etc. My reading of Jurafsky et al.'s data is that, after applying a range of careful controls, there are by and large no lemma effects left. This certainly reflects the spirit of our theory. Still, there are some effects. In particular, there is less coda reduction in partitive *of* than in genitive and complement *of*. Also, determiner *that* never showed vowel reduction as the other *that* lemmas do, and pronominal *that* tends to be longer than the other *thats*. Is there a left-over confounding factor involved? My only hunch is that these differences may relate to the prosodic structure of the word's immediate environment. The data base did not allow for the marking of stress and accent, but it could certainly have been the case in this corpus that the determiner *that* is more often accented, in particular contrastively, than for instance the relative *that*. This needs further testing.

The analysis raises a further theoretical issue worth considering. How does onset latency (the strict empirical domain of the theory in Levelt et al., 1999) relate to articulatory realization, in particular word duration? Is the realization of the articulatory gesture, in particular its duration,

affected by the flow of activation at higher levels of processing? How modular is articulation? Kello et al. (2000) argue against full modularity. In their experiments they used a Stroop task, in which the subject names the color of a printed word. If the word happens to be the name of a different color (e.g. the word GREEN printed in red) color naming latency is substantially slower than when the

word is the name of the color (e.g. the word RED printed in red) or a neutral word (e.g. the word CHAIR printed in red). This is called 'Stroop interference'. In Kello et al.'s experimental results this difference in latencies had no counterpart in the articulatory durations of the color word response ('red'). So far, articulation seemed to be modular with respect to higher level interference. However, when the authors applied a response deadline, requiring a speeded response, they found some evidence for prolonged articulatory duration under Stroop interference. Kello et al. concluded that a modular system can change its architecture to a cascading one, dependent on the specific task demands. I will call this 'restricted modularity'. However, they obtained this effect in only one of their experiments, precisely the one in which 25 % of the data had to be removed for various reasons.

Damian (2001) took up this topic in two carefully controlled experiments. The first one was a picture/word interference task. Here he obtained the usual phonological facilitation effect when the name of the auditory distractor word was phonologically related to the name of the target picture. This facilitation in naming latencies did not 'spill over' to articulatory durations, just as there was no spill over in Kello et al.'s original Stroop experiment. However, even when Damian applied a deadline, which speeded up the responses, again not the slightest effect of phonological relatedness showed up in the response durations. In a second experiment Damian affected response latencies by way of a semantic manipulation. A block of pictures to be named was either homogeneous in semantic category (e.g., all vehicles, or all vegetables) or heterogeneous (a mix of vehicle, vegetable, etc.). Semantic homogeneity leads to substantial interference, i.e., longer response latencies (Kroll and Stewart, 1994; Damian et al. 2001). In agreement with the original Kello et al. data and with the previous picture/word interference data, there was no concomitant effect on articulatory durations. But such an effect was also not obtained when Damian applied a deadline, which led to generally shorter naming latencies. In other words, in carefully controlled experiments, no 'spilling over' or cascading under time pressure could be demonstrated. Thus, so far there is no good evidence against the modularity of

articulation. This, however, may have theoretical repercussions for analyses such as those reported by Jurafsky et al. Assume that a lemma effect does exist. It should show up in response latencies. But if articulation is indeed modular, the lemma effect will not affect articulatory duration. Worse, if Kello et al. were to be right after all, there will not be a spill-over at normal, non-speeded articulation rates. Under both theoretical cases, therefore, a null effect in Jurafsky et al.'s data does not guarantee the absence of a lemma effect. Still, if an even more detailed analysis of these duration data showed a lemma effect, then both theoretical positions (full and restricted modularity) would be in trouble.

Let me now turn to Schiller et al's contribution on syllable priming. Syllable priming has never been a happy topic. In a most careful dissertation, Baumann (1995) showed that, whatever one does experimentally, it is impossible to obtain a syllabic priming effect by way of auditorily presented syllable primes. Of course, one always finds auditory priming, but it is irrelevant whether the prime corresponds exactly to a syllable of the target word. The only thing that matters is the number of phonological segments shared between prime and target. We call this the *segmental priming effect:* the more shared segments, the more effective the prime is. This work was done on Dutch and German and one shouldn't exclude the possibility that specific syllabic priming (beyond segmental priming) would be possible for other languages, for instance syllable timed languages. In their paper Schiller et al. provide convincing evidence against syllable priming in Dutch, English, Spanish, and with a slight hedge, also in French, even though Spanish and French are clearly syllable-timed.

It is theoretically important to observe that the non-existence of syllable priming is in fact predicted by the WEAVER+ + model of phonological encoding (Roelofs 1997, see also Levelt et al., 1999). As mentioned in the introduction, auditory and orthographic priming affect the level of phonological code retrieval. The code is, however, not a syllabified structure. A syllable prime has no special status in code retrieval. Could it affect syllabification? In the theory each auditory or orthographic input segment can prime all related gestural scores in the syllabary. But again, the

syllabic status of the prime is irrelevant. If the first syllable of a target word is of the type CVC, then a corresponding CV prime will prime it partially, a corresponding CVC prime will prime it fully, and a corresponding CVCC prime will prime it fully and the next syllable partially. The total amount of priming, therefore, is simply a function of the number of segments, not of syllable structure. In short, WEAVER++ predicts prime length effects, but no syllable effects.

It is essentially for the same reason that WEAVER++ predicts a number-of-segments effect in phonological word encoding, but no number-of-syllables effect. This was recently challenged by Santiago et al. (2000) on the basis of MacKay's Node Structure Theory (NST). They claimed to have obtained a number-of-syllables effect but no (independent) length effect. However, the authors had not fully controlled for word length in their experiments. A reanalysis by Roelofs (2001) shows that there is only a length effect in the reported data, no independent number-of-syllables effect.

However, the total absence of syllable priming and syllable structure effects in latency measurements of phonological encoding does mean at all that syllables are not essential planning units in speech production. Syllables are among the earliest acquired and most frequently produced motor programs in our repertoire. They are, in fact, so heavily overused, that it would be impossible *not* to store them. Of course, one can construct new ones, but that is an exceedingly rare event. In our theory, phonetic, articulatory syllables are major, ultimate targets of form encoding. It should be added that these gestural scores may well have internal hierarchical structure not unlike the syllabic structures proposed in NST. Motor priming experiments by Sevald, Dell, and Cole (1995) show that such phonetic syllabic structure is independent of phonemic content.

The paper by van Heuven and Haan, finally, moves us beyond phonological word encoding. It addresses an important issue in higher, supra-word level phonological encoding. From the point of view of encoding, the main data reported are the measured sentence melodies of statements versus declarative questions. In these measurements the syntax was kept constant; we are getting a pure

view of intonational differences between the two sentence moods. Clearly and expectedly, the two melodies differ markedly in their boundary tones. But the more interesting finding is that they also differ in other respects. Apparently, speakers give away the mode of their utterance long before they generate the boundary tone. There is, in particular, a difference in the balance of the two pitch accents in the sentence melody. To the best of my knowledge, this phenomenon has not been reported before. The authors argue convincingly that this difference results from an amalgamate of several encoding processes. It is not the case that the speaker decides to make an interrogative statement and then incrementally installs these melodic features. Rather, each feature is installed for its own reason. In particular, if the speaker wants to invite the interlocutor's confirmation that a particular referent was involved in some state of affairs, the speaker will focus that referent by pitch accent. In itself, this has nothing to do with interrogation, but it does cause the characteristic imbalance of pitch accents. It should not be too complicated to elaborate this decompositional approach, and test it in the laboratory.

One advantage of this theory of intonational encoding is that it allows for a limited planning window. Speakers are often under so much time pressure that they cannot afford much 'look ahead' in their conceptual, grammatical and phonological encoding. The architecture of speech encoding must allow for piecemeal, incremental planning (Levelt, 1989). A decompositional stepwise encoding of question intonation relieves the speaker of attentionally loading long-term planning, but still, the outcome will be a natural pitch contour. I am not claiming that speakers always operate at such a minimal look ahead level. When there is no particular time pressure, or when the speaking situation is a more formal one, incremental encoding units can become larger, especially for skilled speakers.

It should not go unnoticed that the main experimental contribution of van Heuven and Haan's paper is their finding that listeners can, by and large, pick up the characteristic pre-boundary pitch cues. However, that moves us beyond phonological encoding.

The studies discussed in this section support a general view of phonological encoding as a multilevel, incremental process. The levels of encoding, from intonational to syllabic, involve dedicated and rather modular operations. Incrementality is achieved by minimizing 'look ahead' at all levels of processing.

Notes

1 Glossary 'Learning to read ... reading to learn' of the National Center to Improve the Tools of Educators. http://ldonline.com/ld_indepth/reading/ltr-cec/ltr7-cec.html

2 Assuming 45 talking minutes a day, two words per second, 1.5 syllable word length.

References

Baars, B. J., Motley, M. T., and MacKay, D
1975 Output editing for lexical status from artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 14,* 382—391.

Baumann, M.
1995 *The production of syllables in connected speech.* Unpublished doctoral dissertation. Nijmegen University.

Crompton, A.
1982 Syllables and segments in speech production. In: A. Cutler (ed.), *Slips of the tongue and language production.* The Hague: Mouton.

Damian, M. F.
2001 Articulatory duration in single word speech production: Evidence for modularity.

Damian, M. F., Vigliocco, G., & Levelt, W. J. M.
2001 Effects of semantic context in the naming of pictures and words. *Cognition, 00,* 00-00

Garrett, M. F.
1975 The analysis of sentence production. In G.H. Bower (ed.). *The psychology of learning and motivation: Vol. 9.* New York: Academic Press.

Kello, C. T., Plaut, D. C, & MacWhinney, B.
2000 The task-dependence of staged vs. cascaded processing: An empirical and computational study of Stroop interference in speech production. *Journal of Experimental Psychology: General, 129,* 340—361.

Kroll, J. & Stewart, E.
1994 Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149-174.

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

Levelt, W. J. M.
1992 Accessing words in speech production. Stages, processes, and representations. *Cognition, 42,* 1-22.

Levelt, W. J. M.
1999 Models of word production. *Trends in Cognitive Sciences, 3,* 223-232.

Levelt, W. J. M. & Wheeldon, L.
1994 Do speakers have access to a mental syllabary? *Cognition, 50,* 239-269.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S.
1999 A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1—38.

Lupker, S. J.
1982 The role of phonetic and orthographic similarity in picture-word interference. *Canadian Journal of Psychology, 36,* 349-367.

Meringer, R. and Mayer, K.
1895 *Versprechen und Verlesen.* Stuttgart: Goschensche Verlag. (Re-issued, with introductory essay by A. Cutler and D. A. Fay (1978). Amsterdam: Benjamins.)

Meyer, A. S.
1990 The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language, 29,* 524 - 545.

Meyer A. S. & Schriefers H.
1991 Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory Cognition, 17,* 1146-1160.

Roelofs, A.
1997 The WEAVER model of word-form encoding in speech production. *Cognition, 64,* 249-284.

Roelofs, A.
1999 Phonological segments and features as planning units in speech production. *Language and Cognitive Processes, 14,* 173-200.

Roelofs, A.
2001 Syllable structure effects turn out to be word length effects: Comments on Santiago et al. (2000). *Language and Cognitive Processes,* 00,00-00

Santiago, J., MacKay, D. G., Palma, A., & Rho, C.
2000 Sequential activation processes in producing words and syllables: Evidence from picture naming. *Language and Cognitive Processes, 15,* 1-44.
Schriefers, H., Meyer, A. S. & Levelt, W. J. M.
1990 Exploring the time course of lexical access in speech production: Picture-word interference studies. *Journal of Memory and Language,* 29,86-102.
Sevald, C. A., Dell, G. S. & Cole, J. S.
1995 Syllable structure in speech production: Are syllables chunks or schemas? *Journal of Memory and Language, 34,* 807-820.
Shattuck-Hufnagel, S.
1979 Speech errors as evidence for a serial-ordering mechanism in sentence production. In W.E. Cooper & E. C. T. Walker (eds.). *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum.