A Prosodic View of Word Form Encoding for Speech Production[*]

Patricia Keating, Phonetics Laboratory, UCLA
Stefanie Shattuck-Hufnagel, Speech Group, Research Laboratory of Electronics, MIT

Note: This paper grew out of an invitation to Pat Keating to summarize and critique the model of phonological encoding presented in Levelt, Roelofs and Meyer (1999) at the Labphon7 meeting in Nijmegen, the Netherlands, in 2000. As a result, and because this model specifically addresses the problems of phonological and phonetic processing which concern us here, we focus largely on it, leaving little opportunity to discuss the advances in other models of the speech production planning process, as developed by e.g. Butterworth (1989), Crompton (1982), Dell (1986), Dell et al. (1997), Ferreira (1993), Fromkin (1971, 1973), Fujimura (1993), Garrett (1975, 1976, 1984), MacKay (1972), Pierrehumbert (in press), Shattuck-Hufnagel (1979, 1992), Stemberger (1985), Vousden et al. (2000), and others.

1. Introduction

In our view (shared with Levelt), one of the most interesting questions in the study of language is how a speaker crosses the boundary between abstract symbolic representations and concrete motor control programs. Given this assumption, we can ask what are the processes that take as input a sentence with its words, and provide as output a quantitative representation which can guide the articulatory motor system in producing a particular utterance of that sentence. An articulated utterance must by definition have concrete patterns of timing and amplitude, as well as an F0 contour and a rhythmic structure. Because these patterns are influenced by, but not fully specified by, the underlying morphosyntactic representation, generating a sentence with its words does not suffice to prepare the utterance for articulation; a futher process is required to generate the structures which will permit computation of the appropriate articulatory patterns. The process of generating the representation which will support an articulatory plan for an utterance, on the basis of the morphosyntactic specification of its underlying sentence and other kinds of information, has come to be called Phonological Encoding (Levelt 1989).

The meaning of the term Phonological Encoding is not entirely obvious to non-psycholinguists, so the first task is to highlight how phonology is part of a code. The term ‘code’ is used in different subfields with different meanings, and the usages that are most likely familiar to phoneticians and phonologists do not correspond to the one intended by psycholinguists modeling speech production. This Phonological Encoding is not precisely equivalent to the Speech Code of Liberman et al. (1967) (an encoding of phonology by speech), nor to the encoding of phonologies by orthographies, nor to the encoding of speech into phonology during speech processing or acquisition (as discussed by Dell 2000 or Plaut & Kello 1998), nor even to the encoding of higher level linguistic structures by phonology (a recent example being Steriade 2000 on paradigm uniformity effects). The psycholinguistic usage is quite subtle, yet simple, compared to these others, if one distinguishes between two uses of the term. The first (the usage in Levelt (1989)) is the larger process of word form encoding, including various aspects we will describe below, while the second (the usage in Levelt, Roelofs and Meyer (1999)) is the process by which an incomplete phonological representation is completed. This second meaning

[*] We would like to thank Bruce Hayes for helpful discussion, and the organizers and audience of LabPhon7.

includes the set of processes that take as input an abstract and somewhat skeletal lexical representation of a word's phonology, and generates a representation which is phonologically complete. Later steps in the planning process which specify the surface phonetic shape that this phonological representation will take have been called Phonetic Encoding. The result of these two processing steps is that the information stored in the lexicon about the forms of the words in a sentence is not the same as the information in the planned sound structure for a specific utterance of that sentence.

Over the past few decades, it has become increasingly clear that these processes must refer to prosody. That is, the timing and amplitude patterns of articulatory gestures in speech are systematically governed not only by contrastive representations of segments and features (at the level of the individual word), but also by suprasegmental aspects of the speaker's plan, which are collectively referred to as sentence-level, utterance-level or phrase-level prosody. Some of the evidence for this view will be discussed below. As we have come to understand more about how phrase-level prosodic constituent structure and prosodic prominence influence the phonetic implementation of utterances, the need to integrate word-level and phrase-level planning in models of speech production has become more urgent. This need was addressed in Levelt's comprehensive 1989 book on the speech production process by focusing one chapter on the generation of phonetic plans for single words, and another on the generation of such plans for connected speech. The conceptual division between single-word and connected-speech processing is further reflected in the 1999 BBS target paper by Levelt, Roelofs and Meyer (LRM99). LRM99 develops the model of single-word planning in greater detail, describing its computer implementation (Roelofs 1992, 1997) along with extensive experimental evaluation. The experiments cited in support of their model use psycholinguistic methods to measure psycholinguistic aspects of language behavior, such as reaction times in primed naming tasks or lexical decision tasks. In addition, the computer model can simulate the effects of experimental manipulations such as priming, and these simulations have provided supporting results that parallel and extend the human behavioral findings. The model provides a detailed account of how speakers might cross the rift between abstract underlying word specifications and a phonetically-specified plan that can guide the articulation of a single word-level element. However, discussion of how this word-level planning process is integrated with prosodic planning for an utterance as a whole is limited to the generation of the word-level prosodic constituent called the Prosodic or Phonological Word (PWd)[1] (Hayes 1989, Nespor and Vogel 1986, Selkirk 1995, Peperkamp 1997, Hall and Kleinhenz 1997). Issues relating to the hierarchy of prosodic structures involving e.g. Phonological Phrases and Intonational Phrases and their concomitants, such as rhythm and intonational contours, are handled by referring to the discussion of planning for connected speech in Levelt's 1989 book (L89); in chapter 10 of that volume we find a description of how higher-level prosodic structure can be built from PWd elements.

The model of single-word planning in LRM99 is considerably more detailed than the L89 version in some respects and more limited in scope in others. It provides an opportunity to evaluate the Max Planck Institute group's approach to single-word production planning in the light of new information that has emerged in the intervening years from prosodic theory, from acoustic and articulatory analysis of surface phonetic variation, and from behavioral experimentation, and also in light of the view of connected speech planning laid out in L89. The

[1] The abbreviation PWd, which we will use throughout, has been used in the literature to refer to both the term Prosodic Word and the term Phonological Word; it appears that the content of these two terms is similar.

model proposed by this group is the most thoroughly worked out in the literature, so we will focus our discussion on it, although the points we wish to highlight are equally relevant to other approaches to modeling speech production. It is our hope that a discussion of these issues, however limited in this format, will encourage more intensive interaction among practitioners of linguistics, acoustics and psycholinguistics, who may have much to learn from a better understanding of each other's theories, methods and critical questions.

In this paper, we first provide brief descriptions of the single-word planning model as described in LRM99 (section 2) and of connected speech planning in L89 (section 3), and discuss problems with these incremental PWd-based models (section 4). Finally, in section 5, we sketch an alternative approach to modeling the integration of word-level planning with phrase-level prosodic planning, which asks the following question: How much of a phrase-level prosodic frame could be built without detailed knowledge of the PWd forms of an utterance, rather than relying on PWds as the basic building block for higher-level structure?

2. Single word production planning

Figure 1 shows LRM99's schematic of the entire model of speech production reproduced from the paper. Note that it starts at the top with concepts and ends at the bottom with sound waves. As noted above, the model postulates a series of processes in the middle which together take the surface syntactic structure of a sentence and its selected words as input, and produce as output the phonetic plan which will govern the articulators as they produce an utterance of that sentence. The three processes are Morphological Encoding, by which all the morphemes needed for a word are accessed, Phonological Encoding (in its narrow sense) and Phonetic Encoding, by which an articulatory plan is formed[2]. Earlier processes shown in the figure, such as conceptual preparation and lexical selection, and later processes of articulation, will not be discussed here, nor will the self-monitoring loop shown at the left. We will focus on the process of Phonological Encoding, and the two processes which bracket it, which collectively we will call Word Form Encoding.

2.1. Where Phonological Encoding fits in the model

In this model, Phonological Encoding is what gets a speaker from the phonology of morphemes to PWds. There are four distinct types of information about a word in the lexicon: semantic and syntactic information (indicated by *lemmas* in the lexicon at the right side of Figure 1)[3], and morphological and phonological information (indicated by *word forms* in the figure). A key feature of the model is that semantic/syntactic information is stored separately from word form information.

[2] The “encoding” terminology replaces the L89 description of Morphological, Phonological and Phonetic “spellout”, reflecting the fact that each of these processes involves more than simple retrieval.

[3] “Lemma” originally meant both the syntax and semantics of a word, but now it is often used more narrowly, to refer to the syntax alone. Figure 1 reflects the earlier usage. The distinction between semantic and syntactic properties is useful in, for example, characterizing function words, which LRM99 describe as “elaborate lemmas but lean lexical concepts”.

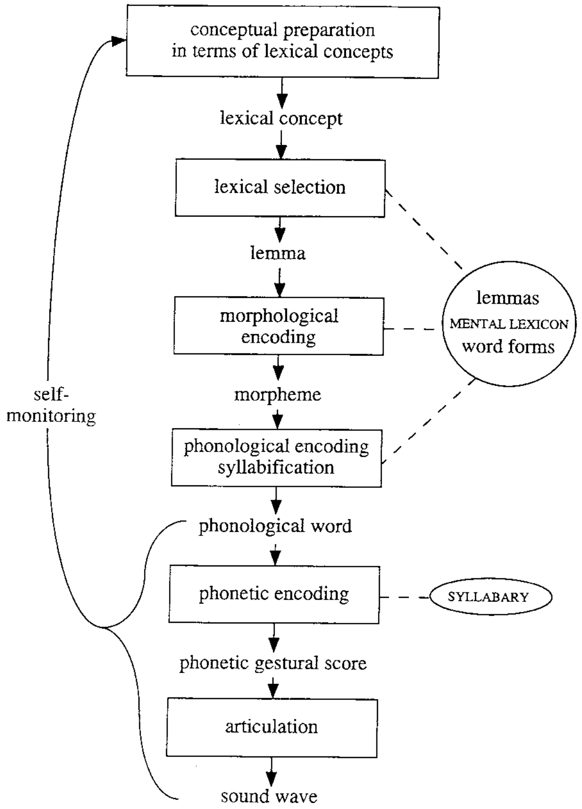

Figure 1. Levelt et al. (1999) model. The boxed elements in this figure refer to processes, the circled elements to stored information, and the unsurrounded elements to products of the processes. See the text for further explication.

Various kinds of evidence have been suggested to support the claim of separate representations of phonological form information, semantic information and syntactic information:

- Word level speech error evidence (L89): Word substitution errors have been classified as two different types, supporting the separate storage of semantic and phonological information. That is, semantic categorical errors (in which the target and error words share membership in a semantic category, such as *month* for *day*, *forward* for *backward*, *French* for *German*) differ from phonological errors (in which the target and errors words are similar in their phonological form, such as *radiator* for *radio*, *similar* for *simpler*). This distinction suggests that the two types of word errors can occur at two different levels of processing, corresponding to retrieval under a lexical concept description vs. from a phonological address.

- Tip of the Tongue (TOT) evidence (L89): This cognitive state provides evidence for separate processing of syntactic and phonological information about a word. Evidence from a number of languages suggests that when speakers are in the TOT state, the lemma, including grammatical gender, is retrieved, but not the (entire) word form (Brown 1991, Vigliocco et al. 1997). This observation suggests a level of processing which involves the morphosyntactic characteristics of a word but not its full phonological form.

- Evidence for order of activation: Further evidence for the separation of phonological and syntactic processing of words comes from experiments showing that lemmas are activated earlier than word forms. For example, studies of event-related patterns of electrical activity in the brain show that Dutch speakers can access the gender of a noun and then stop before accessing its phonological form, but not vice versa (van Turennout et al. 1997, 1998).

- Evidence for constraints on activation: The failure of activated semantic associates of a word to show phonological activation (e.g. presentation of the word *pen* might activate its semantic associate *paper* but there is no sign that *paper*'s phonological form is activated) indicates that a word can be activated without its full phonological form (Levelt et al. 1991). Again, this supports the hypothesis that the two types of information may be stored separately.

Despite this range of supporting evidence, the distinction between lemmas (with their morphosyntactic information) and phonological forms (with their metrical and segmental information) is not uncontroversial, and it is even more controversial whether these two kinds of information can be accessed completely independently of one another (see the commentaries after LRM99).

We turn now to another claim in the LRM99 model: that successive levels of processing can overlap during production planning. As the surface syntactic structure of a sentence is being developed for production and individual words are being selected to fit its lexical concepts, the process of generating an articulatory plan can begin, because this processing is incremental. That is, as soon as enough information is available from one type of processing for an initial portion of the utterance, the next level of processing can begin for that chunk. In the spirit of Fry (1969), and as proposed by Kempen & Hoenkamp (1987), once the surface syntax has been developed for an initial fragment of syntax, word-form encoding can begin for this element while syntactic encoding proceeds for the next section of the message. As a result, the articulatory plan for the beginning of an utterance can be generated and implemented before the entire syntactic plan for the utterance is in place. Such an incremental mechanism is consistent both with the intuition that a speaker doesn't always know how the end of an utterance will go before beginning to utter it, and with some experimental results reported in the literature. It will run into difficulties, however, to the extent that aspects of later portions of an utterance have an effect on the articulation of its beginning. We will return to this issue below.

As noted above, in LRM99's view the generation of an articulatory plan (i.e. the encoding of word form) involves three main components: Morphological, Phonological and Phonetic Encoding. The paper provides an example, shown here as Figure 2, illustrating the steps before and during Phonological Encoding. Correspondences between the strata in this figure and the three encoding processes are somewhat complex, and will be described below.

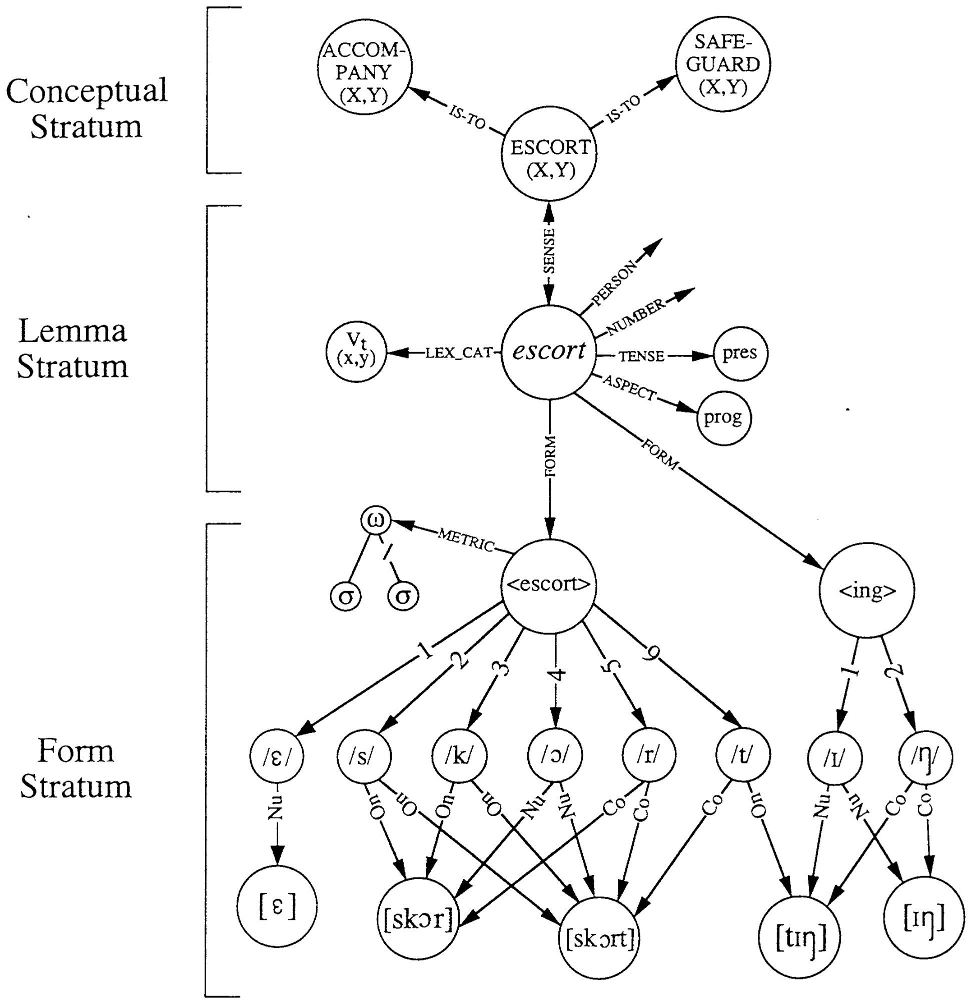

Figure 2. LRM99 Fig. 2, example of encoding for *escorting*.

2.2 Morphological Encoding

The initial steps toward Phonological Encoding are ones that most phoneticians and phonologists don't think about—how, from a lexical concept (Conceptual Stratum in Figure 2), a lemma is found (Lemma Stratum in the figure), and how, from the lemma, its phonological word form is found (Form Stratum in the figure). Each lemma corresponds to a word, at least in languages like English, and specifies (among other things) what kinds of morphemes may be required to create a well-formed version of the word in various contexts. Thus, depending on the lemma's morphosyntactic specifications, at least one morpheme and perhaps more may be required to make the desired form of the word for a particular utterance. For example, in English a verb may require a past or present or present progressive tense marker, a noun may require a plural marker, etc. Retrieving the required morpheme(s) is Morphological Encoding (shown in the top row of the Form Stratum in Figure 2, with nodes between angled brackets). One reason that the step of retrieving the morphemes is necessary before Phonological Encoding can take place is that the specification of morphosyntactic categories in the lemma (such as aspect and tense, shown at the Lemma Stratum in the figure) must be combined to select, for example, the appropriate bound morpheme <ing>. It appears that the morpheme nodes shown between the lemma node and the segmental nodes in the figure represents a stored network of morphemes, separate from the lemma and segment networks. If this is the case, it may have interesting consequences for predictions about error patterns, since misactivation of nodes during retrieval is a prime source of selection errors in such spreading activation models. We will not pursue these implications here, but point out that the ability to generate such predictions is one of the strengths of the model's explicitness[4]. In the example shown in Figure 2, activation of the lexical concept for the action ESCORT has caused competition among various possible lemmas (not shown here), but this competition has ended in the successful selection of the single lemma *escort*. Since *escort* is marked as present-progressive for this utterance, its activation results in the retrieval of two morphemes, <escort> and <ing>. That is, since *escort* is present progressive here, it needs not only the morpheme <escort> but also the morpheme <ing> to be activated in order to complete Morphological Encoding.

2.3 Phonological Encoding

The subprocess of Phonological Encoding (the rest of the Form Stratum in Figure 2) is rather complex in this model, incorporating two kinds of retrieval (metrical and segmental) and a syllabification mechanism. For expository purposes, we have expanded LRM99's figure as Fig. 2a, to show just the portions we discuss below, concerning the Form Stratum.

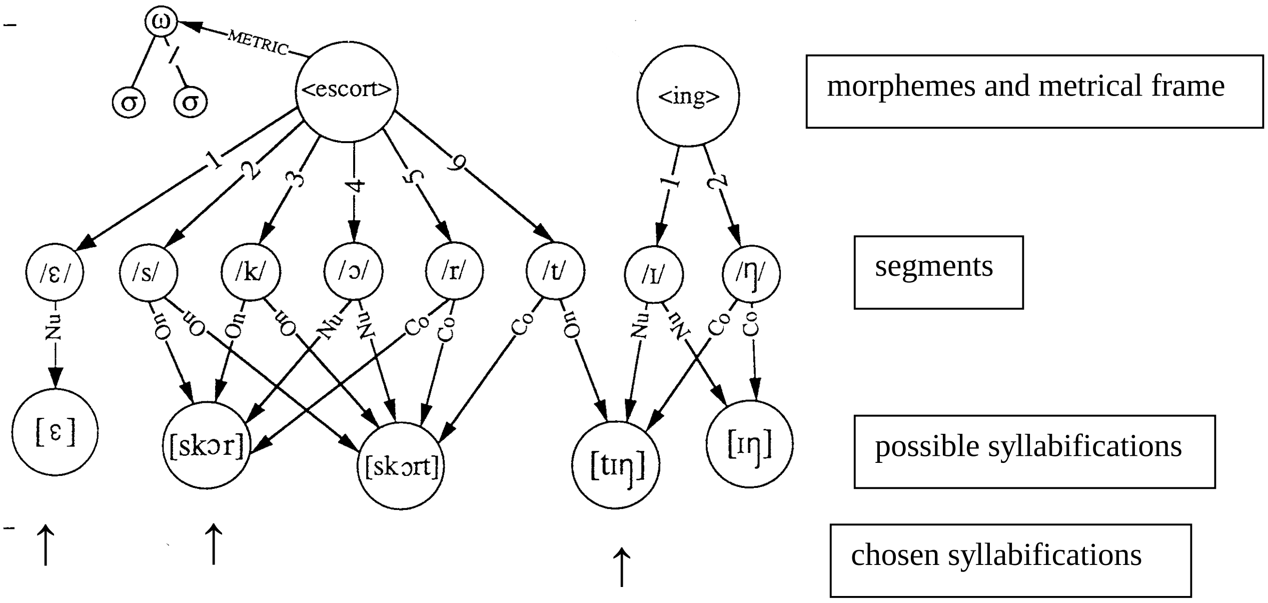

Figure 2a. Expansion of LRM99's figure from above which exemplifies the encoding for *escorting*. This expansion shows only the Form Stratum, with additional labeling.

[4] One of the claims of LRM99 is that for each lexical concept, several lemmas compete for final selection, and that the phonological forms of the competing but unselected lemmas are not phonologically activated; only the word form of the actually-selected (winning) lemma reaches activated status. If two lemmas are somehow both selected, a speech error incorporating phonological elements of the word form information associated with both lemmas can result, such as *symblem* for *symbol* + *emblem*.

Given the appropriate set of morphemes for the lemma, the next step is to activate their phonological forms. These are comprised of two parts: metrical frames, and the segments that fill the frames. Metrical frames specify the number of syllables in a morpheme, and also specify which syllable (*s*) carries the main stress (*S*), e.g. *s S* for the verb form of *escort*, *s s S* for *interrupt,* and *S s s s* for *Abercrombie*. (In the figures above and elsewhere, a slash through a branch in a metrical frame indicates stress.) LRM99's claim that metrical structure and segments are separable parts of phonological form is supported by the fact that when a speaker is in a TOT state, sometimes the metrical frame information is accessed without the complete segmental information that defines the word. In the 1999 version of the model, however, the metrical frame is stored only for metrically irregular (non-default) words; for metrically regular words, a default metrical pattern is computed. This is consistent with the general view that any information that can be expressed as a default specification, or that can be computed from lexical specifications, is not stored[5]. Figure 2a illustrates, in the top row of the Form Stratum, that the iambic metrical pattern of <escort> is not the default English trochee, and so its sS pattern is stored, associated with the morpheme node at the top level of the Form Stratum. In contrast, the suffix <ing> has no metrical frame specified; it is perhaps listed as being simply stressless. When the two morphemes' word forms are retrieved, this sparse metrical information is part of what is retrieved; a later step retrieves the rest, i.e. the ordered string of segments in each morpheme, as shown in the middle row of the Form Stratum.

It appears that in the LRM99 model, speakers represent morphemes (and therefore simplex words) as single nodes, but that the separate morphemes of derived polymorphemic words are stored separately. This means that each of the multiple morphemes for a derived form such as *divinity* or *characteristic* is stored separately and each is linked to its own form information. If this is the case, such complex words must be formed on-line from their component morphemes, a controversial claim. The parallel claim that inflectional morphemes are stored separately from their base forms and that inflected forms must therefore be built on-line is not as controversial.

LRM99 say relatively little about the representation of the segments themselves. They mention briefly arguments by Stemberger and colleagues (e.g. Stemberger 1991) that some segments are underspecified for some phonological features. Although they adopt in the later stage of Phonetic Encoding Browman and Goldstein's (1988, 1992) approach to the representation of gestural scores, they do not explicitly adopt the view that abstract gestures are what characterize contrasts among segments at this phonological level. (In LRM's footnote 6, they argue briefly that differences in surface form among different morphological variants such as *divine/divinity* support a rather abstract representation of phonological form.)

To summarize so far, the elements of stored lexical word form in the LRM99 model are:

- the morphemes (with roots and affixes stored separately; derived forms may be stored as wholes, but definitely not inflected forms)
- the metrical frame of each morpheme (number of syllables, location of main stress), if irregular
- the component phonological segments of each morpheme

[5] Note that as currently implemented, the LRM99 model uses the most common lexical stress pattern and syllable number rather than the linguistically defined default stress pattern to determine unstored metrical frames.

Notice what is *not* part of the stored lexical information in this model: inflected words, syllabification of segments, CV structure, moraic structure, articulatory gestures. Some or all of these representations are developed during the post-retrieval portion of the encoding process, which we will describe below.

Metrical and segmental retrieval from the lexicon are taken to occur in parallel. The metrical frame of a poly-morphemic word is the combination of the metrical frames of the component morphemes – frames which have either been retrieved from the lexicon, or filled in by default. Figure 3 shows the frame that results for <escort> plus <ing>, if a default syllable is inferred for <ing> . This frame is simply the combination of the two component frames, preserving the stress location of the stem. (If derived forms are not stored in the lexicon, however, additional processes will have to follow this concatentation to adjust the resulting frame. These will include such adjustments as changes in the total number of syllables (e.g. for some speakers *piano* as 2 syllables vs. *pianistic* as 4 syllables), in the location of the main stress (e.g. from *pho*- to -*to*- in *photographer* from *photograph*), and in the full vowel reflex of any demoted main stress (e.g. -*lec*- in *electricity* from *electric*), etc, as the result of derivational processes.) Similarly, the segments of a polymorphemic word are the combination of the segments of the component morphemes before any morphophonemic alternations. The segments of the entire word are associated with the metrical frame of the entire word by a process of syllabification.

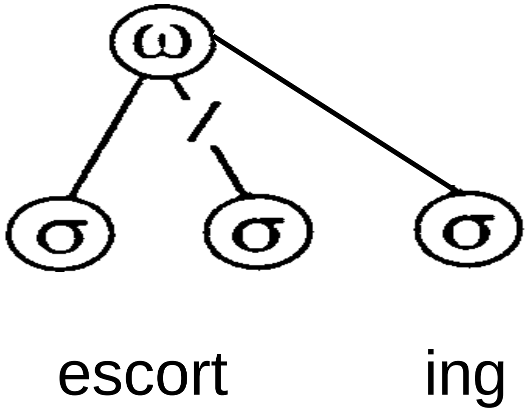

Figure 3. Metrical frame for PWd *escorting*. The slash in the second branch indicates stress on that syllable.

Before describing the syllabification process, however, we must introduce a key idea in this approach to Phonological Encoding, i.e. that the metrical frames to which segmental specifications are associated during syllabification are frames for PWds rather than lexical words. The major motivation for this idea is that monosyllabic function words can sometimes become incorporated with their host lexical words into PWds, which are prosodic constituents. For example, in the LRM99 model, just as *escort* combines with the inflection -*ing* to form the PWd *escorting*, so does *escort* combine with the separate lexical word *us* to form the PWd *escort us*. Experimental evidence from PWd priming studies (Wheeldon and Lahiri 1997) supports the hypothesis that this constituent plays an active role in production planning.

Given this assumption, along with the combined metrical frame and the combined string of segments of a PWd, the string must be syllabified for articulation by fitting each segment into a syllable. Although the details of their implementation are not laid out in this paper, one way of interpreting Figure 2a is that the speaker considers all possible syllables and then selects the optimal path through this set according to syllabification principles. This will ready the string for the selection of syllable-sized articulatory plans from a stored set. Recall that the segments are not syllabified in the lexicon; the metrical frame (with its syllables) is completely separate from the segment information there. That is, the metrical frame specifies the number of syllables, but not the segments which belong in them; the segments are stored separately, and their syllabification is not provided. Only now, in building the PWd for a particular utterance, are segments syllabified. In the Form Stratum of the lexicon, each segment is annotated with its possible positions in the set of possible (i.e. legal) syllables. The choice of actual syllabification from among these possible ones follows general phonological principles (e.g. maximize onsets), as well as language-specific ones (such as ambisyllabicity in English). Crucially, the selection of syllable structures must be made separately for each new utterance, since the choice will vary with the nature of the following material, i.e. whether the next syllable is in the same PWd, and begins with a vowel which could accept a word-final consonant from the preceding word as its onset. Figure 2a shows LRM99's illustration of this with *escorting*. Theoretically, there are several possible three-syllable structures within this word. Under one, the consonant /t/ serves as coda in the second syllable; under another, /t/ serves as the onset of the third syllable (British pronunciation). The syllabification that is activated, on the basis of the structure of the entire PWd including possible inflections and clitics, is shown by the arrows added by us to the bottom of Figure 2a; this syllabification maximizes the onset of the third syllable. This means that the final /t/ of *escort* is now in a syllable associated largely with a different morpheme, *-ing*. For *escort us*, the syllabification process is similar, so that the /t/ syllabifies into a different lexical item for this PWd as well. (This could be called *resyllabification* except that the /t/ had not previously been syllabified; therefore it is referred to as "resyllabification", in quotes.) LRM99 use this "resyllabification" to illustrate why they postulate a process of Phonological Encoding in the first place, in that the *stored* information about form is not the same as the *planned* form that governs the actual pronunciation of the utterance. For example, phonological segments from two different stored lexical entries can find themselves in the same syllable in the PWd structure for a particular utterance. Note that even if lexical words such as *escorting* are precompiled, combinations such as *escort us* surely must be built on-line. In general, a word's pronounced form in an utterance depends on its phrasal context, and thus the planning of its surface phonological form requires an encoding process, since it cannot result from the simple retrieval of the stored phonological form which defines the word and contrasts it with other word forms.[6]

At this point, Phonological Encoding is complete. The phonological segments of the PWd are syllabified into a legal metrical frame, which is now ready for Phonetic Encoding (not shown in Figure 2a).

[6] Another account of surface phonetic variation has recently emerged in the form of episodic theories of the lexicon, which postulate that a speaker stores the range of forms heard and produced for each word, and retrieves the one that is appropriate to the current phrasal context (Pierrehumbert 2001). This approach moves the issue that is described here as an encoding problem into the domain of storage and retrieval.

2.4. Phonetic Encoding

Phonetic Encoding in this model begins with the phonological syllables that emerge from Phonological Encoding. Given a string of phonological syllables, Phonetic Encoding consists of finding their counterparts in a syllabary of stored syllabic gestural scores, or of constructing such scores for syllables which are not stored. The hypothesis is that gestural scores (as proposed by Browman and Goldstein 1990) of high-frequency syllables are stored in a syllabary, while rarer syllables' gestural scores must be constructed on-line. Phonetic Encoding of a whole word then consists of retrieving the gestural scores for the individual syllables of the word and combining them. In this respect it is rather like syllable-based concatenative speech synthesis, and thus faces similar issues of contextually-adjusting gestural scores (see discussion in section 4). These concatenated and contextually-adjusted gestural scores can then be articulated in order to pronounce the word.

This approach neatly solves the problem of whether syllables are frames for holding segments, or organized chunks of segmental material. In this model, both claims are true: the phonological syllables are frames, while the phonetic syllables are chunks.

2.5. Relevance for phonologists and phoneticians

In sum, the issues addressed by Phonological Encoding models are essentially the same as for traditional phonology. All such models try to determine:

- the units of representation
- how much of the representation is already in the lexicon
- how contextual variants are computed

The LRM99 model embodies such answers as:

- Units of representation are (sometimes underspecified) phonological segments.
- Word forms in the lexicon contain no syllables or moras, but include sparse metrical frames which specify number of syllables and location of main stress, for non-default forms only.
- Contextual variants are captured by stored syllable templates.

When LRM99 commit to these choices in their model, it is on the basis of their interpretations of the psycholinguistic literature – sometimes existing studies, sometimes studies carried out specifically to settle such questions, but always some kind of experimental evidence. Phonologists and phoneticians should therefore be especially interested in why these choices have been made. Segments are the basic unit of sound representation in this model because that is LRM99's interpretation of the speech error evidence (Fromkin 1971). The decision in favor of underspecified segments also comes from an interpretation of speech error evidence (Stemberger 1991). The reason for positing that information about the number of syllables in a word is stored in the lexicon is that experiments show implicit priming for this aspect of word form. That is, when the speaker knows that all of the words in the target set to be spoken have the same number of syllables, he or she can initiate the utterance faster than if the words in the target set differ in their number of syllables (cited in LRM99 as Meyer et al. in prep). On the assumption that speed of initiation of an utterance reflects the difficulty of lexical retrieval, implicit priming by shared number of syllables suggests facilitation of lexical access by this shared aspect of word form. This interpretation predicts that words without stored metrical forms, i.e. metrically regular words, won't show this effect[7]. Similarly, the location of non-default main stress can be primed, suggesting it is available in the lexicon. Finally, evidence that syllable-internal structure (such as moras or onset-nucleus-coda organization) is not in the lexicon comes from LRM99's interpretation of both kinds of evidence: speech errors do not unambiguously reflect syllable-internal structure (Shattuck-Hufnagel 1992), and there is only limited evidence that syllable structure can be primed (Meijer 1994, Baumann 1995, Schiller 1997, 1998, though see also Meijer 1996 and Sevald et al. 1995). Thus the empirical support which underlies the proposals for processing steps and representations in this model recommends it to the attention of phonologists and phoneticians grappling with the same issues.

Another aspect of the model that was determined by experimental evidence is of practical importance to phoneticians and phonologists. In the model, homophones are different lemmas with the same form, i.e. they share a single word form in the lexicon. This is illustrated in Levelt et al.'s example of MORE and MOOR (for a dialect in which these two words are pronounced the same) (see Figure 4). These words have very different meanings and therefore separate lemmas, but since they are pronounced the same, the question arises as to whether they have identical but separate word forms, or instead share a single word form. The answer in the model comes from experiments comparing how quickly different words are accessed. In general, low-frequency words are accessed more slowly than high-frequency words. However, low-frequency homophones are an exception to this generalization, in that low-frequency homophones like *moor* with high-frequency partners like *more* are accessed as fast as the high-frequency partners are. That is, it seems that the low-frequency homophone inherits the advantage of its high-frequency partner. This result can be captured in the model if (given some assumptions about how access works) the two words share a single word form, and thus it is the same word form that is accessed in the two cases. The lesson from this example concerns the study of frequency-based phonetic differences. If one wants to study whether higher frequency words are pronounced differently from lower frequency words when they occur in the same context, then homophone pairs seem to be an appealing object of study because they are segmentally matched. However, if such pairs share a single word form, then there is no point in studying their pronunciation differences; there could be no phonetic differences between them, except those induced by the different contexts in which they appear (see Jurafsky, in press).

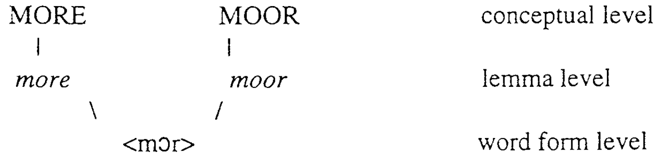

Figure 4. Example of homophones sharing a word form in the lexicon (from Levelt et al. 1999). Note that <mɔr> is not a traditional morpheme but rather the first level of representation of word form information.

[7] However, there is also another interpretation of the priming result: speakers may be able to generate a series of identical metrical frames without retrieving them from the form entry in the lexicon.

3. Levelt 1989 and connected speech

3.1. Model of Phonological Encoding

In his 1989 book *Speaking* (L89), Levelt not only provided an earlier version of the single-word Phonological Encoding model, but also described a model for the creation of phonetic plans for connected speech. This model, also adopted in LRM99, was the first serious attempt to deal with the problem of how speakers plan whole utterances. Levelt emphasized that our understanding of this process was sketchy in the extreme, and it remains true that we have only the most rudimentary knowledge of the full range of phonetic variation and how it arises. Nevertheless, L89 provided a summary of the some of the kinds of information that are required to produce a well-formed utterance of a sentence and an outline of a mechanism for providing this information during the production planning process. We review here the part of this model concerned with Phonological Encoding from Chapter 10 in L89, shown in Figure 5.

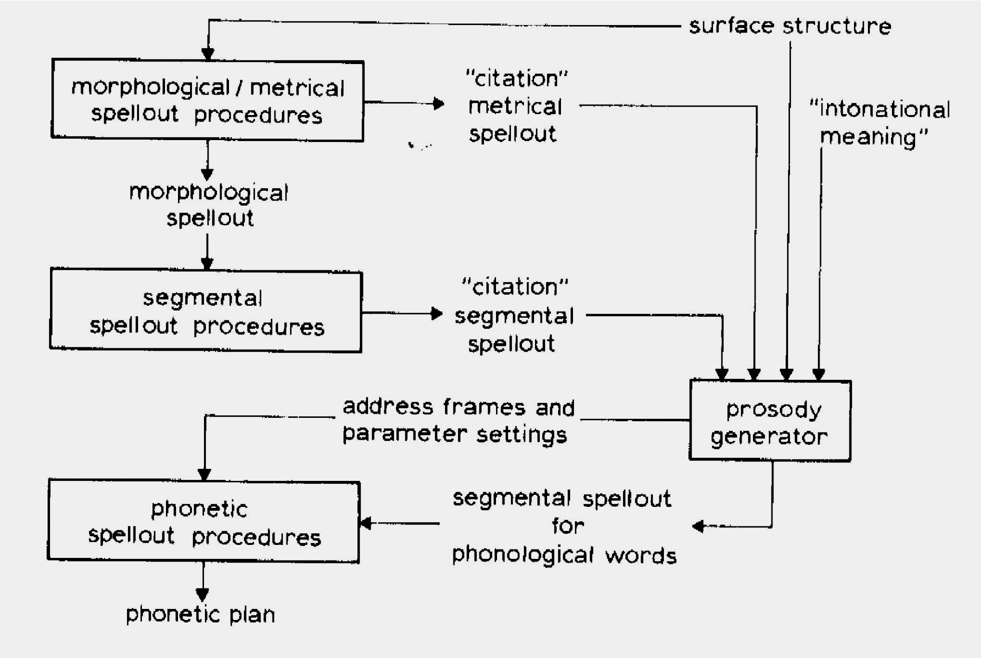
surface structure
morphological / metrical spellout procedures
"citation" metrical spellout
"intonational meaning"
morphological spellout
segmental spellout procedures
"citation" segmental spellout
address frames and parameter settings
prosody generator
phonetic spellout procedures
segmental spellout for phonological words
phonetic plan

Figure 5. The Phonological Encoding model for connected speech from Levelt's 1989 book *Speaking* (Chapter 10, pg. 366).

The left-hand side of the figure shows the components posited in Chapter 9 for single-word processing – morphological, metrical, segmental, and phonetic. Lemmas are used to retrieve a word's morphemes and metrical structure, morphemes are used to access a word's syllables and segments, and segments/clusters of segments are used to address stored phonetic syllable plans. These processing components are all active in connected speech processing as well, but the flow of information between them is somewhat different. That is because the model for generating phonetic plans for connected speech is built around a mechanism called the Prosody Generator, on the right side of the figure, which accomplishes a number of things: 1) it determines the prosodic constituent structure of the utterance, including Phonological Phrases and Intonational Phrases, 2) it determines aspects of phrasal form associated with these constituents in the model: the rhythmic pattern and the intonation contour, and 3) it accounts for the effects of these prosodic characteristics on the phonetic shapes of the segments and syllables. L89 stresses that the plan for a connected utterance is more than “a mere concatenation of word plans. There are, in fact, differences at the level of segments and syllables, at the metrical level, and at the intonation level” (p. 301), and these differences between lexical and connected speech forms are effected by the Prosody Generator.

3.2. Inputs

The Prosody Generator accomplishes these goals on the basis of four pieces of information (the inputs to it that are shown in Figure 5). The first two are aspects of word form information that is being retrieved as activated lemmas become available for word form encoding. This information includes the two separately-stored aspects of a word’s form that are already familiar to us from the single-word encoding model: the metrical frames, and the phonological segments. These are indicated as “citation” forms in the figure because they are here in the form in which they have been retrieved from the lexicon; in L89 there is usually no need for within-word phonological encoding because it is posited that all derived and inflected words are listed in the lexicon. However, the metrical spellout does include any pitch accents on words that were assigned during the building of surface structure. The next input shown is surface syntactic structure (whose construction is the topic of earlier chapters in L89). Surface structure includes not only the obvious information about phrase structure, but also markers for elements that are in focus, and crucially, lemmas, with their morphosyntactic diacritics. The lemmas point to word form information and are input to the left side of the figure, while the syntactic constituency information goes to the Prosody Generator. The final input is the emotional, attitudinal, or rhetorical “intonational meaning”, which will determine the selection of intonation and pitch range, and possibly speaking rate.

3.3. Outputs

The operation of the Prosody Generator on these inputs generates two outputs. The first, “address frames and parameter settings”, refers to syllable templates composed of slots for syllable constituents (onset, nucleus, coda) – the address frames – each “enriched” with values or settings of the prosodic parameters duration, loudness, F0 change, and pauses. The second output is the context-dependent segments of PWds, labeled as onset, nucleus, coda. Lexical words have been restructured into PWds, and their segments have been adjusted accordingly. The segments are then ready to be fit into the slots in the syllable templates.

Together these outputs, now input to phonetic spellout, allow the construction of a phonetic plan for articulating the utterance as a whole (“phonetic spellout”). Once an address frame is filled by segments, it comprises the address for the phonetic plan for that syllable, which can therefore then be retrieved. (That is, there is another input to the phonetic spellout procedures not shown in the figure, namely the set of stored syllable plans.) The prosodic parameter values which have been carried along with the frame get added to the retrieved phonetic syllable. (These parameters do not necessarily affect the segments of a syllable uniformly, and therefore may appear to produce adjustments at the level of the segment. The example given is that the shorter durations of fast speech affect vowels more than consonants, with the result that vowels may appear reduced while consonants do not.) The output of Phonological Encoding, the phonetic plan, is then the input into the Articulator (not shown in Figure 5), which executes the plan.

In order to construct the two outputs shown in the figure, the Prosody Generator constructs other, intermediate, outputs. Most notably, prosodic (called "metrical") structures – not only PWds and utterances but also PPs, IPs -- are constructed by the Prosody Generator, but are not seen as outputs in the figure. Instead these structures influence aspects of the ultimate outputs. Intermediate outputs are constructed from different combinations of the input information.

For example, information from the citation metrical spellouts of the words (which includes pitch accent information in this model) and from the surface structure are combined to give a metrical grid for the sentence. The grid is the basis for computing values along some of the prosodic parameters, such as durations and pauses. The metrical grid information in turn combines with citation segments to allow PWd construction. Only after the segments are put into these PWds, and then adjusted depending on their contexts, does the final segmental output ("segmental spellout for phonological words") obtain. Thus the Prosody Generator affects segmental spellout, because it provides the phrasal/prosodic conditions that affect it. Two examples of this kind of effect in addition to the syllabification effects treated in LRM99 are discussed in L89, but they both are thought to involve syntactic rather than prosodic structure: French liaison as described by Kaisse (1985), and contraction of auxiliaries/clitics. Therefore, the work done by the Prosody Generator with respect to segmental spellout is fairly limited: it uses information about syntactic phrasing that is already in the input, rather than prosodic phrasing computed later. Another case discussed in L89 is assimilation across PWd boundaries, as in *ten books* when the /n/ becomes /m/-like because of the following /b/. This case is special because it operates beyond the domain of a single PWd, and is discussed in the next section.

As another example, surface structure underlies building the grid's constituents, which are phonological phrases (the core metrical unit in this model) and intonational phrases. Metrical information combines with intonational meaning to give rise to the utterance's intonational melody, which consists of a nuclear tone and prenuclear tune drawn from an inventory in the style of Halliday (1967), as described by Crystal (1969). The intonation is still not itself an output, but instead underlies the values assigned to each syllable for the F0 parameter, which is an output. Thus it is clear that the box in Figure 5 labeled "Prosody Generator" contains a number of interesting components which do a lot of work.

3.4. Incremental processing with little lookahead

The L89 model of Phonological Encoding, like other components of the model, operates incrementally and thus ideally requires no lookahead (L89:373 says "very little lookahead"). Levelt takes great pains to show how operations that might seem to require lookahead can be accomodated in an incremental model, thus providing a set of novel and interesting proposals; he also acknowledges some challenges.

An important example is the Rhythm Rule or Early Accent (here, Beat Movement). A word with late prominence, such as *Japanese,* creates a clash with an early prominence on the following word, as in *Japanese Institute*; this clash is resolved by removing the prominence from the offending syllable of the first word, and sometimes making an earlier syllable of that word prominent instead (Horne 1990, Liberman 1975, Selkirk 1984, Shattuck-Hufnagel et al. 1994, Monaghan 1990, Beckman et al. 1990). In order to know whether a clash has occurred or not, he speaker must know the prominence pattern of the upcoming word. The speaker must also know whether the two words are in the same phrase, given that Hayes (1989) and others have claimed that the Rhythm Rule is bounded by the Phonological Phrase. In L89, a syntactic phrase boundary after the first word blocks Beat Movement, but otherwise the Prosody Generator looks ahead to the metrical frame of the next word, to see if Beat Movement should occur. Information about phrase boundaries is given by boundary symbols after phrase-final words, and, conceived of as a property of the word itself, thus hardly seems to count as lookahead. More challenging are cases of iterative Beat Movement, such as *sixteen Japanese Institutes*, because the lookahead encompasses more than the next word. Levelt suggests that such cases are difficult to produce and rare, restricted to more formal speaking styles where speakers plan more carefully. Nonetheless they are a challenge, and the suggested approach to a solution is to posit a larger buffer (and thus longer lookahead) in more careful speech (p. 385).

Other cases of some lookahead include building Phonological Phrases (dealing with final fragments and nonlexical heads), cross-PWd assimilation as in *ten books*, and aspects of intonation (to get the best location for nuclear tone, to incline up to the nucleus, to conjoin a prenuclear pitch accent with the next in a hat pattern). Again some appeal is made to careful speech planning: allowing lookahead in careful speech allows “a more euphonious output, more rhythmic phrasing, and larger melodic lines” (p. 405).

In contrast, progressive final lengthening, which can extend over several syllables and would seem to require looking ahead to anticipate an upcoming phrasal boundary, is treated entirely incrementally as a left-to-right process (p. 390). Each word is longer than the one before, up until a break at the end of an Intonational Phrase.

4. Discussion of these models

As we have seen, the LRM99 model represents a significant advance over earlier, less-explicit models of Word Form Encoding. The model is particularly well worked out for the higher levels of the encoding process, i.e. for Morphological and Phonological Encoding, where the supporting evidence from psycholinguistic experimentation is strong. A key aspect of this model is the integration of the grammatical and lexical information about a sentence into a different kind of structure, i.e. the prosodic structure for the particular utterance of that sentence which the speaker is planning to utter. We are persuaded that this general view is correct. Moreover, the model provides a number of specific advances in the understanding of speech production processes. For example, the fact that the output constituents of Phonological Encoding are prosodic rather than morphosyntactic elements (i.e. PWds rather than lexical words) provides a natural connection to the process of building higher levels of prosodic structure for the utterance. Other strengths of the model, whether new or drawn from the L89 version, include its provision for computing phrase-level prosodic phenomena (such as pitch accent, rhythmic patterns, and intonation contours), its proposal that syllable programs may be specified in terms of articulatory gestures (which appear to be the most appropriate vocabulary for the task of describing systematic phonetic modification and reduction in natural connected speech), and its provision for the computation of phonetic plans from constituents larger or smaller than the syllable. These features highlight the importance of the next step of elaborating precisely how the model accomplishes these tasks. The model also addresses data from priming experiments which shed light on the time course of different aspects of the Word Form Encoding process, it separates semantic, syntactic and phonological information in the lexicon in accord with the empirical evidence, and it introduces the concept of the PWd as the constituent which both forms the planning frame for phonological encoding, and begins the process of generating the prosodic structure of the utterance.

In short, LRM99 builds on and significantly extends the L89 model of single word processing. However, the L89 model of phrase-level planning has not yet been revisited in the same way, and the task of integrating the LRM99 model of single word processing with a model of connected speech processing has been largely left for future work. LRM99 envisioned that using the PWd as their basic unit would make the rest of this task straightforward; however, even in L89 it was not easy to see precisely how this integration could be accomplished, and the greater explicitness of LRM99 has made this difficulty even clearer. In our view the problem lies in the idea of encoding the PWd first, and then doing the prosody later. That is, the problem arises from the fundamental conception, shared by L89 and LRM99, that Word Form Encoding is completely separable from phrase-level processing: that these are two separate things that need to be planned separately and then brought together. This conception is of course a very traditional one, according to which segmental and suprasegmental characteristics can be separated because they control different aspects of the speech signal. However, we believe instead that segmental and suprasegmental characteristics need to be integrated throughout production planning, because prosody is entwined with segmental phonetics. To address these issues, in section 5 we propose a prosody-first model, in which the higher-level prosodic structure of an utterance becomes available as needed during the Word Form Encoding process.

The specific design principles that LRM99 and L89 rely on are not the only principles that are compatible with their general approach of encoding word form by integrating morphosyntactic and prosodic information. In fact, we believe that accumulating evidence about the prosodic factors that influence phonetic variation in word form suggests the value of exploring the different tack which we embark on. We are influenced in this direction by the degree of context-related phonetic variation in word form, the role of higher-level prosodic structure in governing this phonetic variation, and the fact that this variation involves not just the specification of values for traditional prosodic parameters like F0, duration and amplitude of the syllable but also specification of traditionally segmental parameters such as the timing of complex articulatory gestures within a syllable or even a segment. We organize our presentation of the evidence supporting our claim according to two questions about how Word Form Encoding can be dependent on context beyond the PWd. The first is whether the speaker needs to look *ahead* to later portions of the utterance, and the second is whether the speaker needs to look *up* into the higher levels of prosodic constituent structure. The LRM99 answer to both of these questions (with some hedging) is no, following from the basic tenet of incrementality, which postulates that, just as the speaker generates the syntactic structure of a sentence in left-to-right increments, and can begin the further processing of an early increment before later increments are complete, so the speaker also generates the word forms and prosodic structure of an utterance of that sentence incrementally, one PWd at a time, building higher levels of prosodic structure on the PWds as they emerge from the Word Form Encoding process. In contrast, our answer to both of these questions is yes. We also ask a third question: whether the speaker needs to look *inside* the syllable in order to compute the differential effects of context on subsyllabic constituents. Again our answer is yes; the rest of this section explains why we take this view.

4.1. Evidence that speakers look ahead to do Word Form Encoding

There are many observations which indicate that speakers make use of information about upcoming words in order to determine the surface form of the current word. In this section we discuss three such lines of evidence: segmental speech errors across PWd boundaries (4.1.1), stress class resolution across PWd boundaries (4.1.2), and length constraints on prosodic constituents (also 4.1.2).

4.1.1. Speech error interactions across PWd boundaries. Several generations of research in speech errors at the sublexical level have revealed that fragments of words, such as morphemes, segments, strings of segments and occasionally even individual distinctive features can be misordered in the speech planning process, as in *chee kain* for *key chain*, *dretter swying* for *sweater drying*, or *intelephoning stalls* for *installing telephones*. Although such errors occasionally occur within a PWd (e.g. *shiff* for *fish*), the overwhelming majority that have been reported (Fromkin 1971, Shattuck-Hufnagel 1979, 1992) occur between two content words and therefore in at least some cases between PWds. As LRM99 point out, this observation presents a problem for a model in which Word Form Encoding occurs one PWd at a time, because word form information about the lexical items in upcoming PWds is not available while the current PWd is being encoded. Since many interaction errors at the phonological level involve adjacent PWds, L89 proposes a small amount of 'peeking' at word-form information about the next PWd. However, this minimal departure from strict PWd-based incremental processing does not account for the errors that occur across intervening PWds, such as *you getter stop for bas* for *you better stop for gas.* This view also seems to predict more errors within the PWd than across PWds, whereas in fact cross-PWd segmental interaction errors are the norm.

In addition, the encoding of one PWd at a time makes it difficult to account for another aspect of sublexical speech errors: the fact that when misordering errors occur, the elements that are misordered do not wander into random locations in the utterance, but instead occur in each others' target locations. This does not seem particularly surprising in the examples given above, where for example in *chee kains* for *key chains* the displaced target /k/ appears in the next possible syllable onset. But it is more striking in errors where there are intervening onsetless words, such as *the Sicks and the Kneltics* for *the Knicks and the Celtics*. Some mechanism is required to account for the observation that the displaced /n/ from *Knicks* does not appear for example as the onset of the vowel-initial word *and*, but instead appears in the target location for the now-empty onset /s/ of *Celtics*. This pattern suggests the existence of a planning framework of some kind, which specifies that there is no onset consonant for the word *and* but there is an onset consonant for the word *Celtics* (and probably also other relevant information, such as prominence patterns, that contributes to the occurrence of this error). LRM99 argue against the existence of such a framework, on the basis of experimental evidence that prior knowledge of the CV structure of words to be produced does not facilitate production, in contrast to prior knowledge of e.g. number of syllables, which does facilitate production, as predicted by their model. However, other lines of empirical evidence, such as Sevald et al. (1995), do suggest a facilitatory effect of advance knowledge of CV structure. In the face of conflicting experimental evidence on this point, in the approach sketched in section 5 we postulate a planning framework derived from the lexical representation of the words during Phonological Encoding, and larger than a single PWd, which might more easily adapt to this aspect of speech error patterns.

4.1.2. Lookahead in the building of prosodic structure. In the LRM99 model, prosodification, or the building of phrase-level prosodic structure itself, is just as incremental as Word Form Encoding is. Here we consider two difficulties for this view that arise from the necessity for lookahead. As already presented in section 3, L89/LRM99 point out that one aspect of prosodification which requires knowledge about the upcoming PWd is the resolution of stress clash, as in *Japanese Institute*. In order to resolve this clash, the speaker must know the prominence pattern of the upcoming word: regardless of the exact mechanism for clash resolution , there is no way to resolve stress class without looking at more than a single PWd. L89 allows lookahead to the next PWd to account for this. This is a violation of incrementality, but a minimal one. However, iterative cases such as *SIXteen JAPanese INstitutes* provide more of a challenge, and for these cases L89 suggests that the buffer for phonological encoding may be variable in size, i.e. greater than a single PWd. This is an extraordinary proposal to maintain in light of LRM99's focus on the PWd, but it is justified by observations of pitch accenting behavior. Moreover, greater challenges lie ahead, since iterative cases can involve several PWds, as in *MISSissippi LEGislators' ANtique NINEteen TWENty-seven MOtorcars*. A little bit of lookahead can do the job when only one more PWd is involved, but when several PWds are involved, much more lookahead will be needed; and when this idea is taken to its logical conclusion we believe that the model will result in something more like what we will propose in section 5. Moreover, this proposal ensures that word-form information (presumably including segments) will be available across a longer span of the utterance in formal than in casual speech. This suggests with respect to speech errors that more segmental ordering errors, or errors over longer spans, will occur in formal speech than in more relaxed speech, a prediction that seems counterintuitive.

Resolution of stress clash is not the only aspect of prosodic structure building that requires some lookahead. More generally, prominence assignment requires knowledge of more than the current PWd in several senses. For example, there is a tendency to regularize the rhythmic alternation of phrasally strong and weak syllables (Hayes 1984), often termed eurhythmy, and to eliminate stress lapses, i.e. long strings of non-prominent syllables by promoting a non-prominent syllable to greater prominence (Selkirk 1984, Hayes 1984, Nespor and Vogel 1989). Monaghan (1990) points out that in order to establish an alternating rhythm of prominences, whether by deleting clashing ones or inserting others to eliminate lapses, and nevertheless preserve the nuclear (i.e. final) pitch accent of the intonational phrase on the semantically and lexically appropriate syllable, it is necessary to compute this alternation starting from the nuclear pitch accent and moving leftward. Since in many utterances the nuclear accent occurs late, this means that, to the extent that eurhythmy is a tendency in connected speech, information about the prominence patterns later in the utterance must be available in order to compute the prominence pattern for an earlier PWd. Prominence assignment also requires looking up into prosodic structure (about which we will have more to say below): a number of investigations have shown support for Bolinger's (1965) claim that speakers tend to place a pitch accent as early in the first accented word of a new intonational phrase as possible. For example, Shattuck-Hufnagel et al. (1994) documented this phenomenon of Early Accent Placement for FM radio news speech, and Beckman et al. (1990) showed the same pattern for read laboratory speech. The possibility of a variable-sized buffer for phonological encoding could solve this problem, but again it represents a potentially serious retreat from strict incrementality.

An additional line of evidence that speakers have access to some information about upcoming portions of the utterance while building the prosodic structure of earlier parts comes from studies showing that the length of the sentence influences its prosodic constituent structure. For example, Gee and Grosjean (1983) inferred a tendency for phrases (perhaps intonational phrases) to have equal lengths, that is, for the preferred locations for large breaks to come near the halfway points of their sentences. Watson (2002) showed that the length of upcoming portions of the utterance (as well as of earlier parts) influences the speaker's choice of location for Intonational Phrase boundaries in American English. In this spirit, Jun (1993) showed that the number of syllables influences the location of Accentual Phrase boundaries in Korean. Even if this kind of result is limited to read speech, in which speakers are given the overall length of the utterance in advance, it shows that the speech production mechanism is set up to make use of such information. As with the previous cases of stress clash resolution and eurhythmy, the camel’s nose is already under the tent.

4.2. Evidence that speakers look up to higher levels of structure to do Word Form Encoding

Over the past decade or so, increasing evidence has emerged to show that the surface phonetic form of a word in a particular utterance is significantly influenced by the prosodic context in which it occurs, including both prosodic prominence and multiple levels of prosodic constituent structure. As a result, it is reasonable to postulate that the process of Word Form Encoding requires the speaker to have access to this information. We will discuss a sample of the evidence that supports this claim under two headings: evidence for the phonetic effect of prosodic constituent edges, and of prosodic prominences.

4.2.1. Edge effects. Edge effects show that prosodic structure plays an active role in Word Form Encoding, and also reflect the hierarchy of prosodic constituents. The phonetic shape of an individual segment depends not only on its position in its syllable, but also on its position in foot, PWd and larger phrasal constituents. Fougeron (1999) provides a thorough review of studies of the effects of position in prosodic domains, both initial and final, on the realization of individual segments or features. The cases involving initial positions have been called domain-initial strengthening. For example, at LabPhon2 Pierrehumbert & Talkin (1992) showed that /h/ is more consonant-like when it is phrase-initial than when it is phrase-medial, that word-initial vowels are more likely to be laryngealized when they are Intonational-Phrase-initial and/or pitch-accented, and that the Voice Onset Time (VOT) of /t/ is longer phrase-initially. Similarly, at LabPhon6, Keating et al. (in press) presented results from several languages showing that the articulation of a stop consonant’s oral constriction is progressively stronger as the consonant occupies the initial position in progressively higher-level prosodic domains; this strengthening does not affect the entire syllable, but only its first portion. A sample of Keating et al.’s data is shown in Figure 6. This figure, obtained using dynamic electropalatography, shows the pattern of maximum contact between the tongue and the surface of the palate during the articulation of a Korean stop /n/ in different phrasal positions – in Korean, the (post-pausal) Utterance, the Intonational Phrase, the Accentual Phrase, and the Word. As can be seen, the contact is greater when the stop is initial in larger phrasal domains. Such findings provide support for the view that speakers need to know something about the prosodic organization of the utterance above the level of the PWd, in order to complete Word Form Encoding. Reduction processes that operate differently at the edges of prosodic constituents also support this claim. For example, a particle can be reduced when it occurs in the middle of a phrase, as for *up* in e.g. *Look up the word*, but not when it occurs at the end, as in *Look the word up* (Selkirk 1995) even when unaccented. This means that the speaker must know whether there is a boundary after this word, in order to determine the quality of its vowel; it is unlikely that this determination can be made by allomorph selection, in contrast to the case of contracted auxiliaries analyzed in L89:375-380.

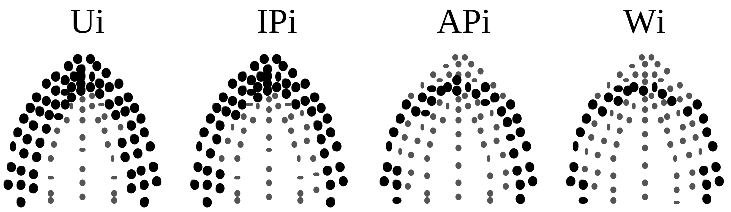

Figure 6. Sample linguopalatal contact data for Korean /n/: each point is an electrode on the surface of the speaker's palate, with larger points indicating contact with the tongue (from Keating et al. in press).

Final lengthening (e.g. Fougeron 1999 review; Cambier-Langeveld (2000)) is perhaps the best-known edge effect. It is challenging to any incremental phonetic model in that it involves not only what appears to be lookahead (because final lengthening may begin before the final word), but also syllable-internal effects (because final lengthening of some segments is greater than that of others). As mentioned in section 3, L89:389-390 suggests an account of this apparent lookahead in terms of progressive (left-to-right) lengthening over the phrase. Yet this mechanism seems implausible in that it predicts that lengthening begins with the second word of a phrase, and also that syllables at the ends of longer phrases should be longer than syllables at the ends of shorter ones. Contra the first prediction, Wightman et al. (1992) found that lengthening did not extend even as far back as the vowel preceding the final syllable.

A final example of edge effects is the pattern of constituent-edge glottalization in American English. For left edges, Dilley et al. (1996) report that reduced word-initial vowels are significantly more likely to be glottalized at the onset of a Full Intonational Phrase than an Intermediate Intonational Phrase. For right edges, Epstein (2002) showed that non-modal phonation is associated with Low boundary tones, but not with low phonetic pitch in general. Redi and Shattuck-Hufnagel (2001) report that final creak is more likely at the end of a Full Intonational Phrase than at the end of an Intermediate Intonational Phrase, a constituent which is lower on the hierarchy. If the creak is a planned variation in voice quality, rather than an automatic consequence of another planned event such as lowered F0 or reduced air flow in these locations, then this pattern provides another piece of evidence for the speaker's need to know something about the higher-level prosodic structure of the utterance in order to carry out Word Form Encoding.

4.2.2 Prosodic prominence effects. Speakers also need to know something about the phrase-level prominences or pitch accents in an utterance in order to complete Word Form Encoding. Since, as we have seen, pitch accent patterns are not solely determined at the PWd level, this means that some aspects of phrase-level prosody, namely prominence patterns, must already be determined when PWd encoding takes place. The effects of prominence on word form have been clearly established; Fougeron (1999) provides a thorough review of such studies. For example, de Jong (1995) suggested that under prominence, segments have more extreme oral articulations that enhance their distinctive characteristics. Edwards et al. (1991) report phonetic differences between syllables which are lengthened due to nuclear pitch accent prominence, and those which are lengthened due to a following prosodic boundary. More recently, Cho (2001) compared the articulation of segments at edges of prosodic domains vs. under prominence. Under accent, segment articulations are more extreme, longer, and faster; at boundaries, segment articulations are more open, longer, and slower. Thus these effects are not the same, which indicates that speakers take into account both aspects of prosodic structure, prominences and edges, in phonetic encoding. Voice quality is similarly affected by prominences as well as boundaries. Pierrehumbert and Talkin (1992) showed that a word-initial vowel is more likely to be glottalized if it is in a pitch accented syllable. Dilley et al. (1996) showed a similar pattern for a wider variety of pitch accent types, and the wide variation in the rate of word-initial vowel glottalization among their five radio news speakers suggests that this parameter may be under voluntary control. Epstein (2002) showed that accented words, regardless of pitch accent type, are like phrase-initial words in that both have a tenser voice quality, and that this relation is not due to positional effects on F0. Like prosodic constituent edge effects, these effects of prosodic prominence on word form, combined with the dependence of prosodic prominence on overall phrasal prosody, show that speakers must know about prosodic aspects of upcoming portions of an utterance in order to complete the Word Form Encoding of the current PWd.

4.3. Evidence that speakers look up and/or ahead to do Word Form Encoding

The distinction between looking up and looking ahead is not always clear, and indeed many, perhaps most, processes arguably involve both. In this section we consider several processes that might involve both looking up and looking ahead. How individual cases might best be classified into our categories is of course not the point here, since our categories are for expositional convenience only; the crucial point is that of looking beyond the current PWd in some way.

4.3.1. Phonological processes dependent on prosodic structure beyond the PWd. There are many cases in the phonological literature suggesting that speakers make use of information about later PWds, as well as about their prosodic positions, during the Word Form Encoding of earlier words. The thrust of this literature on 'phrasal phonology', which goes back at least to Selkirk (1984) and Nespor and Vogel (1986), is that segmental phonological rules can be conditioned by prosodic domains, in the sense that they apply only within particular domains, or only across particular boundaries. Post-lexical, phrasal phonological (i.e. not phonetic)[8] operations mentioned by LRM99 and/or L89 include syllabification of coda consonants (discussed extensively in LRM99), French liaison (L89:367), and English r-insertion (L89: 302); Hayes and Lahiri (1991) describe two assimilatory rules of Bengali that apply within Phonological Phrases; Selkirk (1986) provides other examples of phonological rules that are part of the "sentence phonology". Here we mention two further examples of phrasal (across PWd) operations. The first is Korean nasalization assimilation as described by Jun (1993). In Korean, certain final consonants become nasals when before nasals in the same Intonational Phrase. Jun carried out a phonetic experiment that established that the derived nasals are phonetically just like underlying nasals, that is, the neutralization is complete. Thus, on our view as well as Jun's,

[8] We take phonological rules to be those which change the value of a distinctive feature, adjust stress or insert/delete a segment, resulting in a different word form; in contrast, processes which change the timing or shape of articulatory configurations that implement the distinctive features of phonological segments we view as phonetic.

this phrasal alternation could not be due to prosodically-conditioned phonetic encoding; instead it must be due to prosodically-conditioned phonological encoding. This prosodic conditioning involves access to information about both the prosodic structure of the utterance (because the rule is blocked by an Intonational Phrase boundary) and the upcoming segment (because the rule refers to a following nasal). The second example is Chaga leftward spread of lexical H* as described by McHugh (1990: 56-58). In Chaga, a Bantu language with both lexical tone and word-level accent, a H* accent spreads leftward up to a preceding H tone within a Phonological Phrase. Crucially, a Chaga PP can be several words long; thus the tone of the first word in a PP can depend on whether the last word in the PP bears a H*. Thus lookahead in phonological encoding must extend well beyond the next PWd in Chaga, and again must refer to both the prosodic position (in this case, re the Phonological Phrase) and the lexical form information (in this case, lexical accent) of that upcoming PWd.

The one instance of lookahead which is incorporated into LRM99 is the evidence that a morpheme-final consonant can syllabify into the onset of a vowel-initial syllable across the boundary between two lexical items ("resyllabification", described in section 2.3 above). As evidence for this claim, LRM99 cite the onset-like pronunciation (in British English) of the final /t/ of *escort* not only in inflected forms like *escorting*, but also in phrases like *escort us* which involve more than one lexical word. Although this particular example probably does not hold for American English (see Hayes 1989 for discussion of differential behavior of affixes vs. clitics), resyllabification across word boundaries is attested in other languages and must be accounted for in any model of encoding. As we have seen, this evidence is taken as support for the proposal that segments retrieved from the lexicon are serially ordered into a PWd-sized metrical frame, that this frame is sometimes built from the metrical frames of several lexical words, and that because syllabification occurs during this segment-to-frame association process, syllabification across lexical boundaries within a PWd is possible. Thus, LRM99 incorporates lookahead at the phonological shape of a later word in one limited context: when the later word is a clitic which forms a PWd with the preceding lexical word. That is, this "resyllabification", operating within PWds and bounded by PWd boundaries, is exactly and only the sort of prosody-sensitive phonological process that the model can express. Yet there is an entire typology of prosody-sensitive phonology, as laid out in the theory of the prosodic hierarchy: there is a wide variety rules operating within and bounded by all of the domains of the hierarchy. Resyllabification within a PWd is just one possibility in this typology, but the model is set up to accomodate only this kind of case. Thus it is not a coincidence that LRM99 focus on this process rather than those bounded by other prosodic domains. In our view, "resyllabification"does not provide support for the LRM99 model, but instead only underscores the PWd-specific view of prosodic phonology the model incorporates.

4.3.2. Assimilation across PWd boundaries. As noted in section 3, L89 describes the apparent assimilation of the /n/ in *ten books* to a labial /m/, under the influence of the following labial /b/, and recognizes that such assimilations, involving two PWds, require limited lookahead. (Obviously, assimilations within a single PWd are straightforwardly handled without lookahead. LRM99 would presumably consider assimilations such as partially palatalized /s/ in *miss you* (Zue and Shattuck-Hufnagel 1979) as occurring within a single PWd.) Whether lookahead between PWds involves looking up, looking ahead, or both, seems to us to depend on the account of assimilation itself that is assumed. On the one hand, a traditional view of assimilation (feature-changing, feature-spreading) or a traditional view of local coarticulation, in which prosodic domain limitations are not considered, requires looking ahead to an upcoming segment. On the other hand, if assimilation is simply the result of the overlap of invariant gestures, and is sensitive only to prosodic structure (e.g. Byrd et al. 2000) and not to the content of the gestures, then only looking up in the prosodic structure is needed (that is, the occurrence of the overlap is insensitive to the gestures; of course the assimilatory result of the overlap will depend on which gestures overlap). Other views of cross-word assimilation, that involve both kinds of information, are doubtless possible. For example, Cho (2001) describes the dependence of cross-word vowel-to-vowel coarticulation on higher-level prosodic structure. The point to be made here is that some sort of information beyond the single PWd – up, ahead, or both – is required on all such accounts. As far as we can tell, the lookahead that is proposed in L89/LRM99, is only for the purpose of determining whether to incorporate the next syntactic word into the current PWd, and thus occurs only within the syntactic representation; it thus cannot provide the form information needed for cross-PWd assimilation.

As we have seen in sections 4.1 through 4.3, speakers must look up into higher levels of prosodic structure and ahead to the next PWd and beyond, in order to compute the surface form of words in an utterance. L89 (and by extension LRM99) acknowledges these requirements, and handles them by a series of small departures from the strict principles of incremental Word Form Encoding of one PWd at a time and subsequent construction of higher levels of prosodic structure. These moves raise the question of why, if broad look-ahead and look-up is possible in certain speaking circumstances, it is not available for the planning of speech to be produced in all circumstances.

4.4. Evidence that speakers look inside to do Word Form Encoding

Another principle of the model is its reliance on the retrieval of syllable-sized articulatory gesture scores to account for the effects of syllable position as well as coarticulation among segments within the syllable. The model departs from the principle of phonetic encoding via retrieval of gestural scores from a precompiled syllabary in several ways. First, the syllabary is proposed only for high-frequency syllables, so that an on-line process for computing gestural scores from segments and features must be posited for low-frequency syllables. Second, as acknowledged by L89/LRM99, adjustments to the syllables will be required after syllables are concatenated into PWds, to produce such effects as vowel-to-vowel coarticulation. Such cross-syllable adjustments will also be required after PWds are concatenated into phrases, since as mentioned above, vowel-to-vowel coarticulation crosses PWd and even larger boundaries. At this point it becomes apparent that the retrieval of gestural scores is the beginning rather than the end of the phonetic encoding process. Moreover, the adjustments that are required suggest that the syllable may not always be the most useful constituent for which to store precompiled gestural scores. For example, in American English, many intervocalic consonants are ambisyllabic, as in *upper* or *Ada*. The clear intuition described by LRM99 for the rightward syllable affiliation of the /t/ in *escorting*, presumably from speakers of British English, is not so convincingly available to speakers of American English. In words with ambisyllabic consonants, the assignment of consonants to preceding or following syllables is uncertain, and the hypothesis that it is syllable-sized articulatory units that are concatenated into PWds may need re-examination, even though the concept of precompiled abstract gestural scores of some kind is an attractive idea.

A final example of departure from the principle of phonetic encoding via syllable score lookup is L89's treatment of phrase-final lengthening, as noted in section 4.2.1. In a good-faith effort to reconcile this well-known effect with incremental processing, a mechanism is proposed for beginning the lengthening process after the first word, and increasingly lengthening each word until the Intonational Phrase is completed. Although this proposal avoids the necessity for lookahead, it seems to make a number of counterintuitive predictions, e.g. that all non-initial words in a long phrase will be somewhat lengthened, and that the final word or words of a longer phrase will be more lengthened than the same words in a shorter phrase.

In our view, a wide variety of phonetic phenomena which appear to require manipulation of elements smaller than the syllable also suggest that when the syllable-sized articulatory scores are retrieved, the work of phonetic encoding has just begun. For example, Berkovits (1993) has shown (for Hebrew) that in carrying out utterance-final lengthening, speakers lengthen the onset consonant of the final syllable less than the nuclear vowel, which is in turn lengthened less than the coda consonant. Similar effects have been reported for English by Wightman et al. (1992) and Turk (1999). Even the individual gestures of a complex phonological segment have been shown to be independently timed. For example, Sproat and Fujimura (1983) studied the timing of the tongue-tip gesture of a constituent-final English /l/ with respect to the tongue-body gesture, and reported that the tip gesture was delayed in proportion to the duration of the preceding vowel, which presumably reflects the level of the constituent in the hierarchy. Gick (1999) has reported similar effects for the gestures of English /r/. The L89 model includes a post-syllable-selection phonetic adjustment procedure which presumably provides for the contextual adjustment of timing and amplitude of the articulatory gestures of the syllable scores, so this kind of sub-syllabic processing is not incompatible with the spirit of the model. However, L89 is at some pains to describe the operation of even this post-prosody-generation procedure in incremental terms which do not involve lookahead, and as we have seen there are reasons to wonder whether this can work. Moreover, a number of phenomena that have been documented since 1989 expand the work that must be done by a post-prosodic sub-syllabic mechanism for phonetic adjustment, i.e. a mechanism that cannot easily operate on the syllable as a whole, but appears to require access to individual gestural components. For example, left edge effects of the sort reviewed in section 4.2.1 appear to affect only the first segment of a French syllable, as for example the /k/ in a /kl/ cluster, or the vowel in a vowel-initial word (Fougeron 1998). As the operation of this aspect of Word Form Encoding is described more explicitly in future work, we suspect that it will come to resemble the sort of ‘prosody first’ approach described below in Section 5.

Another set of Word Form Encoding phenomena which the model does not yet address involves suprasyllabic phonetic effects, i.e. effects on whole words. For example, Wright (in press) reports the hyperarticulation of ‘hard words’, e.g. those which are low in frequency and/or in a high-density lexical neighborhood. Similarly, Jurafsky et al. (in press) and others have described phonetic effects of word frequency and predictability. The current L89/LRM99 model does not explicitly address these recently-documented effects, but again it may be possible to do so by marking encoded PWds for later adjustment in a post-syllable-retrieval process.

In sum, it appears that the integration of prosodic parameter values with the syllable-sized gestural scores will require not only look ahead and look up, but also the manipulation of subsyllabic structures and a mechanism for computing phonetic effects on larger constituents such as whole words. While the addition of a processing stage that can accomplish these things is not incompatible with the L89/LRM99 model, and in fact is hinted at by some of their discussion, it expands the scope of Word Form Encoding to include substantial post-prosodic adjustments of the forms of PWds in light of their larger structural and segmental context. It remains to be seen how this requirement can be reconciled with the incremental approach which is central to the single-word-utterance-based view. The point we wish to emphasize is that, even after the pre-compiled gestural scores for the syllables of a PWd have been retrieved, the speaker is still quite far from articulatory implementation—i.e. a good deal more processing is required to specify the systematic differences that we observe both with adjacent context and with higher-level structure, as well as with non-grammatical factors such as speaking rate and style or register. The lines of evidence summarized above show that, in carrying out Word Form Encoding, particularly its phonetic aspects but also its phonological processes, speakers make use of information about later constituents and higher-level constituents than the current PWd. How can the look-ahead and look-up requirements be satisfied, while preserving the substantial insights embodied in the LRM99 model? The following section sketches one view of how this might be accomplished.

5. Building prosody before Phonological Encoding

An alternative to the local, incremental construction of prosodic specifications upwards from the information available at each PWd is the initial construction of a higher-level prosodic representation of the entire utterance, with subsequent integration of word-form information into this larger structure. That is, the speaker may construct at least some higher-level prosodic representation without reference to word forms, providing a locus for those aspects of prosodic processing that do not require word-form information. As more information about both word forms and non-grammatical factors becomes available, restructuring of this initial default representation may then occur. Many different types of information about words are needed for planning the production of an utterance, including such things as their number and serial order, whether or not a word will receive phonological content in this utterance, metrical information such as the number of syllables and stress pattern, the number and serial order of contrastive phonemic segments within a word, and the articulatory instructions for realizing these contrasts. Our hypothesis is that not all of this information about the words of an utterance is required at the same time; as a result, various kinds of information can be retrieved separately, as needed for further processing. This view of 'build prosody first, then retrieve segments' contrasts with the 'retrieve segments first, then organize them into prosodic constituents' view.

5.1. Motivation and general principles

What is the motivation for computing higher levels of prosodic structure for an utterance early on in the production planning process, before Word Form Encoding instead of after? The previous section presented compelling evidence that all aspects of Word Form Encoding, including Phonetic Encoding, must refer to prosodic structure. This prosodic structure therefore must have already been built. But in order to build this prosodic structure, we need to take account of a number of factors, of which morphosyntax is just one. Our hypothesis is that the

.

initial prosodic structure is derived directly from the syntax, and then restructured on the basis of non-syntactic information. Some aspects of this prosodic restructuring can be carried out in the absence of word form information; others require at least some information about word form such as number of syllables and stress pattern; and still others require full knowledge of the phonological segments of the words. Our general approach, then, is to break down the process of computing the prosodic structure for an utterance into two stages. The first pass creates default prosodic constituents based on the syntax/semantics (that is, on non-word-form information). The second pass is influenced by word-form and prosodic information. The approach we sketch here is based on three assumptions about the time course of prosodic planning and retrieval of this word form information:

a) different aspects of prosodic restructuring require different kinds of word form information;

b) word form information becomes available in stages, as LRM99 suggest; and

c) those aspects of prosodic restructuring which can be carried out with minimal word information are carried out early in the planning process, before complete word-form information is available; restructuring into the final prosodic representation, which requires complete word form information, is carried out later.

By a prosodic representation, we mean a hierarchical grouping of words into higher level constituents, such as phrases of different levels, along with indications of the relative prominence of different constituents, intonational accents marking prominences, and any tonal markings of constituent boundaries. Without committing ourselves to any particular view of what the constituents must be, or whether they can be recursively nested, we will use here, for illustration, a prosodic hierarchy with multiple levels: an Utterance consists of a string of one or more Intonational Phrases (IP), which in turn consist of a string of one or more Phonological/Intermediate (Intonational) Phrases (PP/IntermIP), which are made up of a string of one or more PWds. We adopt this simplified hierarchy here for the purposes of illustration; for example, we do not include a separate constituent to account for the difference in phonetic behavior between PWds like *editing* and larger groupings like *edit it* in American, as discussed in Hayes (1989) and Shattuck-Hufnagel (forthcoming). The reader is referred to Shattuck-Hufnagel and Turk (1996) for a comparison of the full constituent hierarchies posited by various prosodic theories.

What is the relation between such a prosodic representation of an utterance and the surface syntactic structure of the underlying sentence? Syntax is clearly a significant factor in determining the prosody, but it is not the only factor. Like others working in the prosodic hierarchy tradition, we envision that the prosodic representation is the locus for integrating morphosyntactic influences with non-grammatical influences such as speaking rate, information structure, affect etc. On this view, these disparate factors exercise their influence on surface phonetic form by influencing the prosodic structure of an utterance. In this discussion, however, we will focus on the grammatical factors, i.e. on the interaction of the evolving prosodic representation with two aspects of morphosyntax: surface syntax, and word form. We will have less to say about how the effects of non-grammatical factors are incorporated into the planning process. Our basic assumption is this: during Phonological Encoding, the process of constructing the prosodic representation takes advantage of information in stages. This means building a skeletal default prosody at first, and restructuring this initial prosodic representation as more information becomes available, especially word form information.[9]

Because we propose restructuring in part on the basis of word form information, our prosodic representation develops in phases. The first, default, representation is based only on the syntax and thus is closely related to it. It is nonetheless a prosodic representation because its constituents are not necessarily identical to any syntactic constituents. The prosodic constituents produced by the first pass of prosody-building are tested in a second pass against form-based constraints, which may trigger restructuring of the constituents. As a result of restructuring, they become less closely tied to the syntax, and therefore will appear more purely prosodic. Our notion of restructuring differs somewhat from that in some of the prosodic literature, in that our restructuring takes one kind of prosodic representation (syntax-based) and turns it into another kind (based on many factors). A constituent which has satisfied the form-based constraints is a true prosodic constituent. In fact, we take it to be a constituent of a surface prosodic hierarchy such as provided by the ToBI transcription system. In general, we will assume a final prosodic representation of constituents and tones consistent with the ToBI transcription system (Silverman et al. 1992), which is in turn consistent with Pierrehumbert (1980), Beckman and Pierrehumbert (1986), and Price et al. (1991). That is, the result of filtering syntax-based prosodic structures through form-based constraints is a set of surface prosodic structures as in ToBI. Thus, to anticipate our discussion below, we take PPs to be syntactically-defined first-pass prosodic structures, but (at least in English) IntermIPs to be the surface constituents that result from checking PPs against the relevant form-based constraints and restructuring them accordingly.

We make two simplifying assumptions: first, that all of syntax is already available, and thus the default prosodic structure can be built all at once rather than incrementally; and second, that there is a series of restructurings moving gradually closer to the surface form. These simplifications may or may not be true, but we adopt them as an expository strategy. If prosodic structure is built incrementally, our claim is that the increments must be large enough to account for the facts of phonological and phonetic segmental sensitivity to prosodic structure. Our account here will have a top-down flavor (see papers in Inkelas and Zec (1990) for some top-down approaches), but our commitment is only to ‘prosody first’. Similarly, there are alternatives to the step-by-step processing that we describe, but we want to make clear that some processing can be done as soon as each new piece of information becomes available.

[9] The possibility of even more radical restructuring of syntax-derived prosodic constituents is raised by a number of proposals for deletion or movement of multi-word strings based on prosodic structure. For example, Booij (1985) has shown that in Dutch, the material that can be deleted in one of a pair of parallel conjoined structures (as in *the number of positive- and negative-tending reports*) is a prosodic constituent. In order to know whether the first instance can be deleted, the speaker must know that this constituent will be repeated in a later parallel structure. Another example: Selkirk (2001) has argued that the criterion for heaviness in heavy NP shift is computed in prosodic rather than in morphosyntactic units; Cohen (2001) has reported experimental evidence to support this view. These proposals suggest the possibility that at least some of the phenomena traditionally assigned to the syntactic component of the grammar may be more usefully envisioned as the result of restructuring of the default prosody derived from surface syntax. The approach to prosodic restructuring described in this section leaves room for such a mechanism, although the types of restructuring included there are less radical in nature and do not involve deletion or movement of word or morpheme strings.

5.2. Deriving an utterance: prosody first

5.2.1. Constructing the default prosody from syntax. A morphosyntactic representation contains some information about the words which will appear in the utterance, i.e. their approximate number (based on the number of lemmas), their serial order, their location in syntactic structure, and the fact that they are not empty elements. It contains no information about word form, i.e. no morphological or phonological form information. How much prosody can we build from such a representation? Here we assume the availability of the syntactic tree whose terminal nodes are lemmas but not word forms. These trees are at least partially marked for prominence, for reasons having to do with semantics or information structure. (See L89, section 5.2.2, for how a particular lemma comes to bear the prominence associated with a larger constituent.) Generally these non-form-based prominences will be sparsely distributed, since their function is to call attention to particular important information in the utterance. In Figure 7 we show a complete structure along these lines for the sentence *The puppies chased those hippopotamuses*, in which the two nouns have been marked as prominent by asterisks, though other words could equally well have been made prominent. These prominences will be associated with pitch accents. Since it is likely that semantic and discourse-level information determine also the types of pitch accent as well as the words which will carry them, this selection could also be made here. In contrast, decisions about boundary tone type must await phrasal restructuring, although some indication of constraints imposed on this choice by speech-act implications (e.g. question vs. statement) may be added to the representation at this point. In addition, decisions about which syllable in the word will carry the accent must await further processing, since the syllable structure of the words is not yet available.

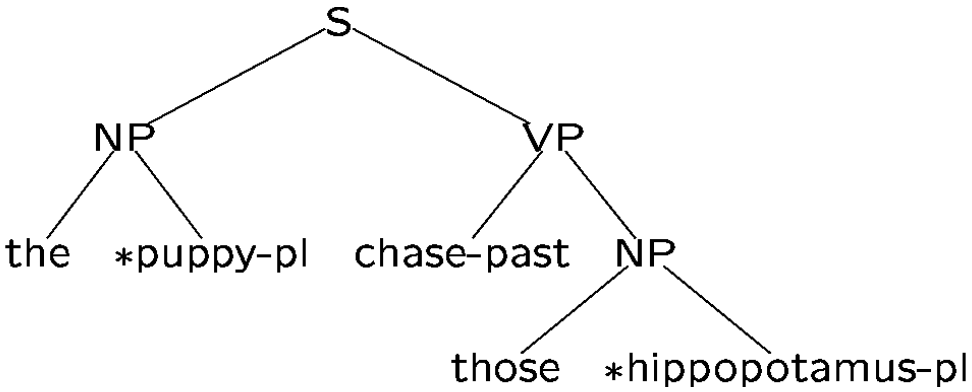

Figure 7. Schematic syntactic structure for example sentence, *The puppies chased those hippopotamuses*, with prominent lemmas (as determined by syntax, semantics, pragmatics) marked with askterisks. This representation contains no prosodic structure or word form information.

On the basis of this representation, we construct an initial, default, prosodic skeleton. Following L89, we know that the Utterance will consist of at least one IP, one PP and one PWd. How many more of each of these constituents will be needed here? First we consider the IP. It is generally agreed that certain syntactic constructions, such as parentheticals, tags, and non-restrictive relative clauses, will occupy their own IPs, and Nespor & Vogel (1986) note that root sentences delimit IPs. Our sentence is a single simplex clause, so it will have only the minimum: one Intonational Phrase. The default PP construction is also based on syntactic constituency, as described by e.g. Selkirk 1986, Nespor and Vogel 1986, Hayes 1989: form a PP from the left edge of XP up to X, optionally adding a non-branching complement to the right. In our sentence, this will give two PPs, corresponding to subject and predicate. Finally, a default PWd structure is copied from the lemmas at the bottom of the surface syntactic tree, assuming one PWd for each syntactic word that will have phonological content (which is all of them in this utterance). For the minimal prosodic structure we are employing here, this completes construction of the initial syntax-based prosodic representation of the utterance, shown in Figure 8.

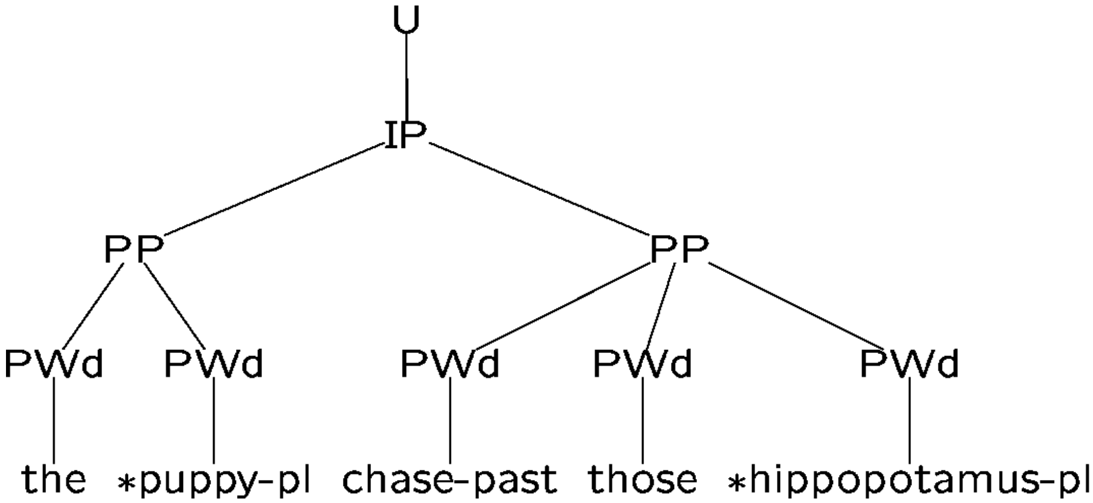

Figure 8. Default prosodic structure for *The puppies chased those hippopotamuses*, derived from the syntactic structure shown in the previous figure.

In such a syntax-derived default prosodic structure, we have less word form information than is assumed in LRM99. There it is assumed that all the phonological form information about a word that is stored in the lexicon, including metrical frames and segments, is available for the construction of a PWd, and presumably it remains available after the PWd is constructed. (L89's Ch. 10 Prosody Generator works this way: higher levels of prosodic structure are built on PWds, and PWds are built with complete word form information.) In contrast, the representation in our Figure 8 contains *no* word form information, because at this point none has yet been retrieved. Our approach will be to restructure the default prosody on the basis of additional information, information about the forms of the words. To what extent can we proceed incrementally in accessing form information from the lexicon, accessing it only as we have exhausted the possibilities for building prosodic structure, and even then, accessing only enough to proceed a bit further? In other words, how much prosodic processing can be accomplished with how little word form information?

5.2.2. Restructuring the default prosody on the basis of word form information. As we noted above, the syntactic tree tells us roughly how many words there are in the sentence because the terminal nodes of the tree are word lemmas. The LRM99 model incorporates the notion of a default metrical frame for English, the trochee. Logically speaking, we could construct a default metrical structure for the sentence as a whole by simply plugging in a trochaic metrical frame for each (non-empty) terminal node in the tree, and we could then use this default metrical structure as the basis for a first pass of form-based restructuring. Although we considered this possibility, we cannot identify any work that such a move would do for us, and so we leave it aside here. Instead, we consider restructuring based on information taken from the lexicon. Although much research remains to be done in this area, it seems clear that some aspects of prosodic phrasing depend on information that is not available in the prominence-marked syntactic tree in Figure 7. L89 (section 10.2.3) discusses additional factors influencing IP breaks, such as desire for intelligibility, rate, length, and prominence relations. Nespor and Vogel (1986: 7.2) cite length, rate, and style, in addition to prominence, as factors influencing their restructuring process. Jun (1993: Ch. 5) cites rate, focus, phonological weight (word length), and semantic weight.

It seems generally agreed that length is a particularly important factor. Jun (1993: Ch. 5) showed clearly that length matters for the Korean Accentual Phrase: in an experiment where the number of words potentially forming an Accentual Phrase was controlled, but the number of syllables in those words was varied, Accentual Phrases preferentially had five or fewer syllables. If combining words into one phrase would result in more syllables, two phrases were preferred. Thus she showed not only that length matters, but that at least in this particular case, length is counted in terms of number of syllables, not number of words. However, given a lack of other such studies, we do not know whether the relevant metric of length is the same across prosodic domains, or across languages. Therefore we will consider the length metric separately in each section that follows.

5.2.2.1. Intonational Phrase: Nespor and Vogel (1986) proposed that the length metric for IPs is essentially phonetic duration, e.g. the combined effect of length and speaking rate. We will assume, for simplicity in the current speculation, that restructuring of IPs proceeds on the basis of the number of words, but not of their lengths in syllables. Although it is not clear that number of words is the length factor that is important in determining IP structure, information about the approximate number of words is already available from the surface syntactic tree, and thus provides a coarse-grained metric of phrase length that requires no word form information. It is thus like rate, style, or other non-grammatical, but non-word-form, factors influencing phrasing; but at the same time it could be regarded as an aspect of form – phrase form.

Taking number of words as our metric, then, if the number of words is too great, or if the speaker desires shorter IPs for some other reason, the default IP can be broken into two or more IPs: [*The mayor of Chicago*] [*won their support*] (from Selkirk 1984). Similarly, if the speaker desires longer IPs, or if the number of words in two successive is is very small, they may be combined into one IP: [*Who me?*]. Finally, the speaker may choose among several possibilities for location of IP boundaries, as in [*Sesame Street is brought to you by*] [*the Children's Television Network*], vs. [*Sesame Street is brought to you*] [*by the Children's Television Network*] (from Ray Jackendoff, p.c.). The likely locations for these restructured phrase boundaries are constrained by the surface syntax, but the speaker may choose to overcome these constraints.[10]

In other words, the default IP structure can be tested against any form-based constraints, such as length, and can also be revised according to other, non-grammatical requirements of the speaker's communicative and expressive intent. In the case of the IP, we posit no word-form constraints. We assume that an utterance with five words, as in our example utterance, can happily comprise a single IP. Restructuring is optional for our example; we assume it does not

[10] It may be necessary to provide a mechanism for additional IP restructuring after the metrical frames of the words have been retrieved, as when the single lexical word *fantastic* is produced as two separate intonational phrases, *Fan-Tastic*!

apply. Thus the IP determined by default from the syntax has satisfied the form-based requirements for a fully prosodic constituent and is now a Full Intonational Phrase (FullIP) in the ToBI transcription system.

At this point, with the restructuring of IPs complete (vacuously for our example), it is possible to assign the tonal elements associated with the boundaries of FullIPs, i.e. the Boundary Tones, to the edges of the FullIPs. The information required to select among these tones is not fully understood; however, it seems likely that these decisions are based on dialogue structure and information structure constraints, rather than on word form information, which is not yet available. Therefore we posit that FullIP Boundary Tones are assigned at this point (just one for our single FullIP). Figure 9 shows the result of processing of the IP, with L% representing the Boundary Tone.

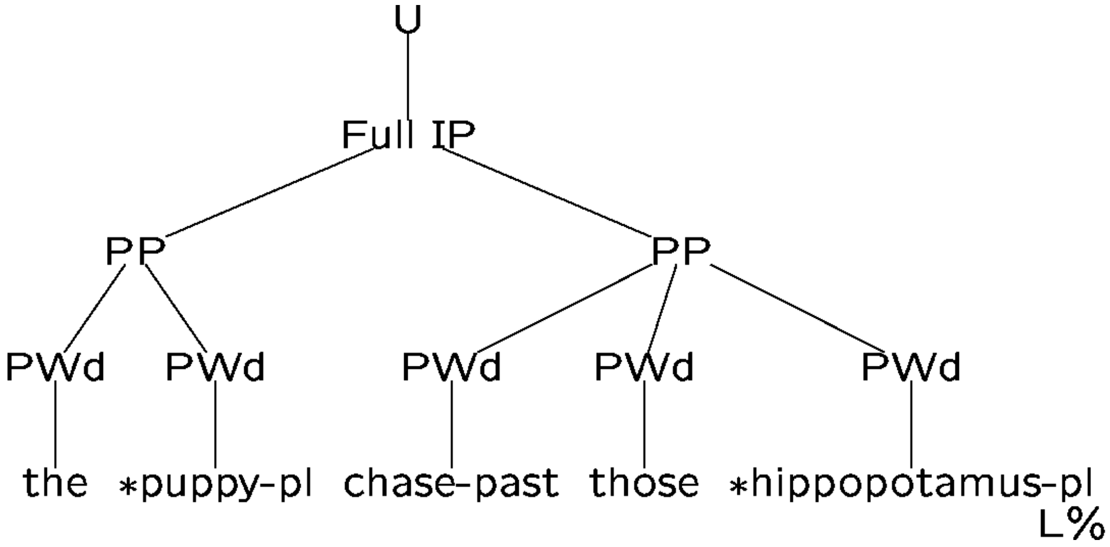

Figure 9. Interim prosodic representation of *The puppies chased those hippopotamuses,* with default prosodic structure (shown in figure 8) restructured at the IP level. L% is a low Boundary Tone.

5.2.2.2. Phonological Phrase. The first-pass syntax-based construction for our utterance resulted in two PPs. On the second pass, these two phrases are checked against the factors that influence phrasing at this level, of which we will discuss only length. One assumption in the literature has been that length constraints operate in terms of the number of immediate constituents of the PP, e.g. the number of Clitic Groups (Nespor and Vogel 1986) or of PWds (Inkelas and Zec 1996). However, as noted above, Jun (1993) showed that the small phrase of Korean, the Accentual Phrase, is sensitive to the number of syllables, rather than the number of words. Let us then assume that, to evaluate the need for PP restructuring, we need to know the number of syllables associated with each lemma. Therefore we now retrieve, from the lexicon, the metrical frame for each morpheme/word corresponding to each lemma. The number of syllables is all we need at this point, but since we adopt LRM99's proposal that the syllable count is bound up in the metrical frame with the stress pattern, we will retrieve the stress pattern too (indicated in the frames by a slash through a syllable branch). If a word has a default stress pattern, and consequently no stored metrical frame, then we could specify a trochaic metrical frame. We plug the metrical frames into their locations, i.e. into the default PWds which were provided by the original lemmas, which have now been organized into IP and PP structures.

Note that we are not retrieving the segments, because we don't need them yet; retrieving frames without segments is possible because they are stored separately in the lexicon, as proposed by LRM99. Adding the newly retrieved word form information, i.e. the metrical frame for each word including its syllable count, to our tree from Figure 9 is shown in Figure 10 below. (We use a question mark to indicate each specified affix, which may or may not turn out to involve an additional syllable.) The terminal elements of this representation still do not contain the phonological segments of the words. Nor are the individual components within syllable structure represented yet, since these are not stored in the lexicon as part of the metrical frame, in the LRM99 proposal we are adopting. These structures must be added in subsequent processing.

As noted above, the syllable count of affixes is not yet known, either because they have no metrical frames, or because the count is subject to phonological adjustment. Thus our composite syllable count may not be quite correct. For example, we probably don't yet know whether *chase*, once it gets inflected for past tense, will turn out to take a non-syllabic allomorph /t/ or /d/, or the syllabic allomorph /əd/. Similar questions arise for *puppies* and *hippopotamuses*. Since there are as yet no sounds associated with the default PWds and their metrical frames, and since the choice among these allomorphs depends on the nature of the word-final segment, we cannot tell whether an additional syllable will be required. L89 suggests that high-frequency inflected forms are stored in the lexicon, but *chased* and *puppies* may not fall into that category, and *hippopotamuses* certainly will not. As another example, in a language which resolves vowel sequences by deletion and/or glide formation, the final syllable count of such words will be one fewer than the starting number. We will assume that such discrepancies will not matter to prosodic structure, though this is an empirical question.

With that assumption, now we can do the form-based checking of our PPs. There are a couple of different ways in which length constraints can be considered. One is whether existing PPs should be combined because they are short. In our example, the first PP *(the puppies*) has few enough syllables that it could combine with another PP to form a single phrase, but suppose that the second PP has too many for their combination (10-plus syllables) to be felicitous. This factor thus favors leaving the default structure of two PPs. For other sentences, e.g. *He sat on the chair next to the fireplace*, with a monosyllabic pronoun subject and a complex predicate, restructuring would be more likely. (As for IPs, other aspects of the speaker's intention may override both the default syntax-based structure and the form-based constraints, as when the speaker chooses to produce each word as a separate constituent to indicate annoyance or impatience with the listener.) Another way in which PP length can be considered is whether the two PPs are currently optimally divided. We have noted that utterances may tend to divide into IPs of equal lengths, and the same could well be true of smaller phrases. Yet the two PPs in our example are of rather different lengths; therefore we will restructure them into two PPs more evenly divided, with the PWd of the verb *chase-past* moving into the first PP. Thus, as a result of length considerations, the prosodic constituency of the utterance moves further away from its syntactic constituency. (Although in our example this result possibly depends on using syllable count as the relevant length metric, the same point could be made for other metrics with other examples; see e.g. Nespor and Vogel (1986).)

In addition to length constraints, there are distributional prominence constraints: there should be at least one pitch accent per phrase. This constraint is met by the default structure, as each default PP contained one prominent word whose prominence was determined on the basis of syntactic/semantic information. Thus no adjustments to the PPs are necessary on this basis; assume that none apply. The PP now corresponds to the Intermediate Intonational Phrase

(IntermIP) in the ToBI transcription system, and perhaps to the Accentual Phrase in certain languages, e.g. Korean.

It might seem that at this point, with the metrical frames available, phrase-level prominences should be mapped to the stressed syllables of the words (which are defined in the metrical frames), rather than to the lemmas. But even though that would work for the words in this example, in general we do not have final stress patterns at this point: because we haven't conjoined the metrical frames of the various morphemes, we haven't made any stress adjustments yet. For example, if derived words like *animation* from *animate* are constructed on the fly during production planning, the final association of prominences to syllables within such words must await completion of those derivations, since reassignment of main lexical stress may be required. We can, however, assign a Phrase Accent (here shown as L-) to a default location near the end of each IntermIP; the final decision about where it will be realized will depend on pitch accent assignment to syllables (see below.). Figure 10 below shows the result of processing of the PPs.

5.2.2.3. Prosodic Word and below. Finally, how many PWds will each PP/IntermIP contain? The crucial form information needed for this step is whether the function words, such as the determiners in our example, can cliticize to an adjacent host. If a word form is stressless, then it can cliticize. In particular, in the example at hand, *the* can cliticize if it is a stressless determiner, although it will not cliticize in some circumstances, e.g. if it is contrastively prominent, as in *THE puppies chased those hippopotamuses, not THREE puppies.* In contrast, *those* will not cliticize, because its metrical frame contains a stress. This information is already at hand, and requires no access to the segmental composition of any words; indeed, L89 proposes that some cliticization is already indicated in the surface syntactic structure, before any form information has been retrieved. Thus we obtain four PWds, and our prosodic structure, restructured on the basis of word form information, looks like Figure 11.

What about prosodic structure below the level of the PWd? In many prosodic hierarchies, between the segment and the PWd are not only the syllable, but also the foot and the mora. Our complete prosodic tree will contain all prosodic constituents, intonational tones, and prominences, whatever they may be.

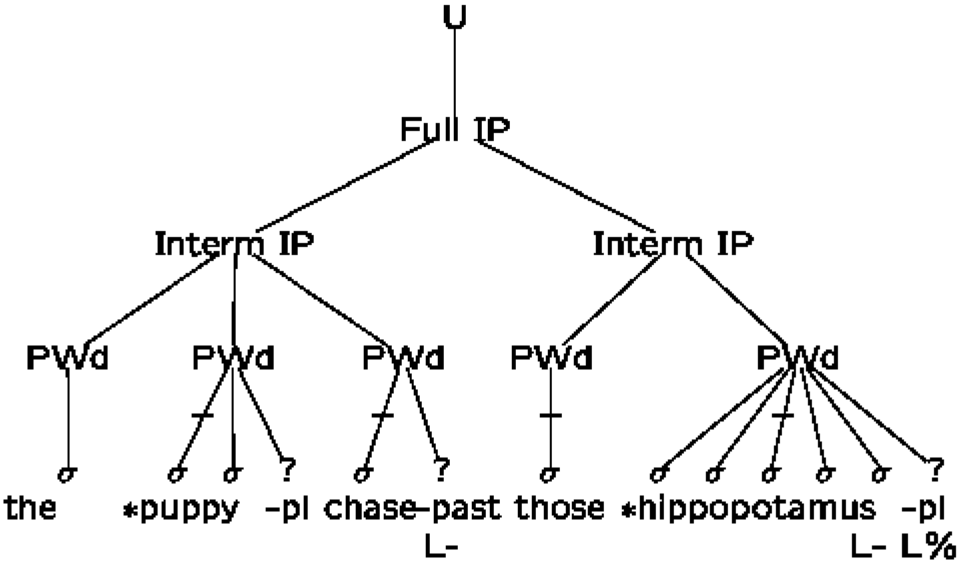

Figure 10. Interim prosodic representation of *The puppies chased those hippopotamuses,* with metrical frames added to lemmas, and with default prosodic structure (shown in figure 8) restructured through the PP level. L- is a low Phrase Accent.

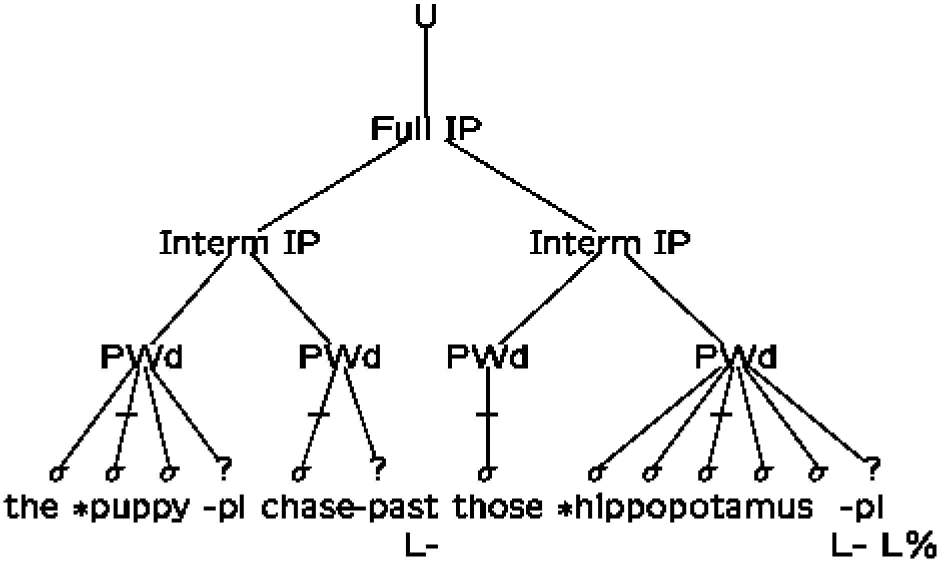

Figure 11. Interim prosodic representation of *The puppies chased those hippopotamuses*, with default prosodic structure (shown in figure 8) restructured through the PWd level.

5.2.3. Phonological Encoding. With the restructuring of prosodic constituents and prominences complete, it is now possible to carry out the processes of Phonological Encoding, i.e. of serially ordering the sound segments of the words, and adapting their phonological shape to the prosodic context. The segments of a word are retrieved from the lexicon, and mapped into the metrical frames of the words in the utterance, which were retrieved earlier. (For simplicity, in our example we follow LRM99 in assigning consonants to syllable onsets wherever possible, even in pre-stress positions where they are arguably ambisyllabic or syllable-final (e.g. the second /p/ in *puppies*).) It is during this process that segment-level sound ordering errors, such as exchanges, may occur. Imagine for example that having built the metrical frame for the entire utterance, the speaker then retrieves the segments for all of the lexical items at once, but maps them into the frame left-to-right. This provides a mechanism for (a) the availability of a downstream segment to be mis-mapped into earlier locations, (b) a frame to maintain the location where the downstream segment should have occurred, so that it can receive the displaced segment from the earlier location, and (c) compatibility with the evidence for left-to-right phonological mapping extensively cited in LRM99.

With word-final segments in place, we can determine the forms of inflections like past (e.g. /t/ in *chased*) and plural (e.g. /z/ in *puppies*), and obtain the final surface syllable structure of the utterance (e.g. that the plural adds a syllable in *hippopotamuses* but not in *puppies*). It is likely that additional restructuring processes take place once the syllable structure of each PWd and the syllable status of each affix is known. For example, we envision that if a canonical syllable structure has been constructed, then empty slots in this syllable will be deleted, making resyllabification of a final consonant possible in some circumstances if the following syllable begins with a vowel. It also may be at this point that phonological consequences of combining clitics with their hosts are computed.

Once the segments are in place, they can be adjusted as appropriate according to their positions in not only the word-sized metrical frames, but in every domain above them. Crucially, the position of every segment in every prosodic constituent can be determined locally by scanning vertically up the tree. As each segment is inserted into the prosodic structure, its encoding (i.e. syllabification, determination of prosodic-context-dependent phonological variation, and possibly restructuring below the level of the PWd) will depend on its position as determined by this vertical scan. Indeed, as described in section 4, much phonological processing could not happen before this point, as it is only now that all the prosodic domains are known. Our example sentence offers no obvious processing of this sort, but, for example, if there is a redundant phonological feature [spread glottis] indicating aspiration, this would be the point at which it would be assigned to the pre-stress /p/ in *puppies* and *hippopotamuses*.

Finally, because all the form information is at hand, we can compute those aspects of form-based restructuring such as structure- and rhythm-governed prominence restructuring, described in section 4. We presume that any such prominence restructuring operations can be carried out in the absence of information about the phonological segments of the words, and as such should be able to occur before segmental phonological encoding. However, we do need to know which syllables carry lexical stress and therefore are potential docking sites for pitch accents, and which syllables are not. This is not known until the PWds themselves, or at least their metrical frames, are built during phonological encoding. Speakers realize prominences (a property of words) by first attaching pitch accents to the main stressed syllables of those words; then any necessary restructuring takes place. Thus we propose that prominence restructuring occurs here, a relatively late operation. In our example, there is no stress clash to resolve, and the tendency for an early and a late pitch accent is already satisfied. Perhaps a pitch accent could be added to break up the long span of unaccented syllables, but the intermediate location is on the function word *those*, which is awkward to accent; and while *chased* is accentable, its location would not be eurhythmic. Therefore we assume no prominence restructuring in this example. Thus in our example, there are two pitch accents, on the first syllable of *puppies* and the third syllable of *hippopotamuses*. Figure 12 shows the results of Phonological Encoding.

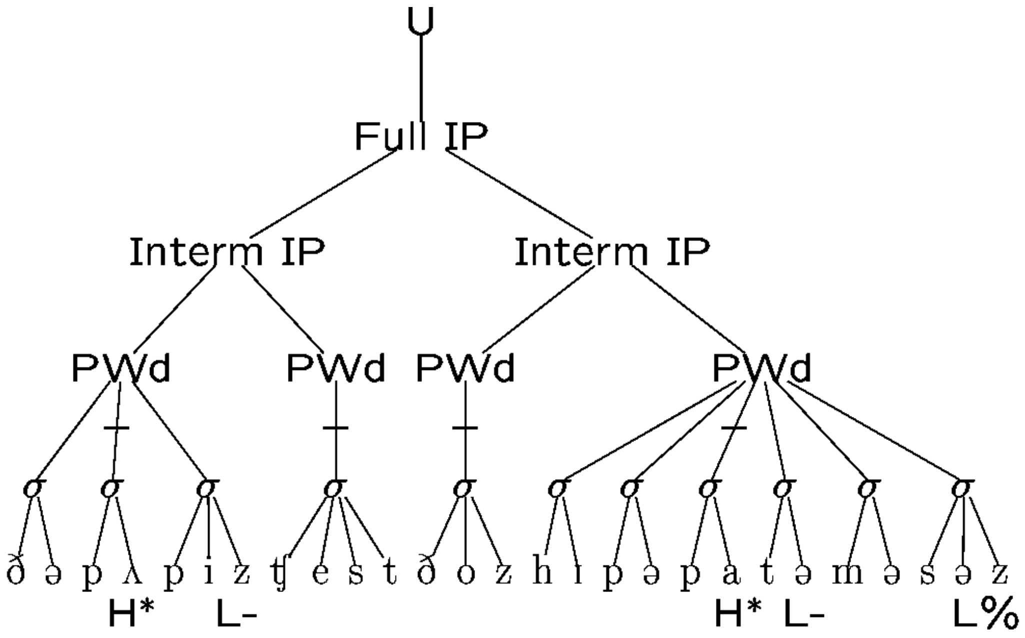

Figure 12. Final (though partial – some domains are omitted) prosodic representation of *The puppies chased those hippopotamuses*, with segments added to metrical frames, and pitch accents associated with stressed syllables. H* is a H-tone pitch accent.

5.2.4. Phonetic encoding. As we have stressed in our discussion so far, the result of Phonological Encoding is still rather far from speech. There is still much work to do, and prosodic structure will play an important role in that work, just as it did for Phonological Encoding. Recall that Phonetic Encoding in LRM99 is usually the retrieval of syllable-sized motor plans from a stored set of gestural scores, and in L89 it is also their realization with the traditional prosodic parameters duration, F0 change, and acoustic amplitude, appropriate for the phrase. In contrast, as earlier sections have made clear, we think of Phonetic Encoding not only as these traditional computations, but also as the shaping of each phonetic feature (or gesture) by its context, both segmental and prosodic.

Consider, as a partial example, the articulatory planning for *the puppies* from our sample sentence, which following LRM99 we have given as a single PWd. Because this begins the utterance, its initial consonant is initial in an Utterance and every smaller prosodic domain; in contrast it is final in its PWd but in no higher domains; in addition, *puppies* bears a lexical stress and a pitch accent on its first syllable. As a result of being in a single PWd, there will be substantial articulatory overlap between the clitic and the host, giving the clitic a short vowel, and vowel-to-vowel interactions should be strong. As a result of being in utterance-initial position, the manner of the /ð/ in *the* is likely to be strengthened to a stop articulation. As a result of being at the end of the PWd, the last syllable of *puppies* will be lengthened somewhat. As a result of the pitch accent on *puppies*, the stop /p/ in the first syllable will be more closely articulated, and the vowel following it will be more open and have a tenser voice quality. As a result of the lexical stress and the pitch accent, that same /p/ will also have a larger glottal opening and thus more aspiration. On the other hand, the word *puppies* is probably not low-frequency or difficult enough to require any special hyperarticulation to enhance its intelligibility, though the sentence as a whole is odd enough that it might.

We have no detailed proposals to make here about how this Phonetic Encoding should be achieved. Following LRM99, it would involve a gestural syllabary for high frequency syllables, and construction of low frequency syllable scores (see section 2.4 above). However, another possibility, which takes both prosodic and segmental context into account, is Keating's (1990a) window model, as developed by Cohn (1990), Keating (1996), and independently by Guenther (e.g. Guenther 1995). Window models posit ranges, rather than fixed points, as the targets of articulatory movements. Segmental interactions, even long-distance ones such as vowel-to-vowel coarticulation across consonants, follow from differences in the ranges attributed to different targets. A large range of possible values allows more influences from context, and a sequence of large ranges allows influences across a large distance. At the same time, lexical, prosodic, and discourse effects can be modeled as hypo- or hyper-articulations, by expanding or shrinking the ranges. Prosodic effects would be local to an edge or a prominence, lexical effects would be local to a word, while discourse effects could be more global.

On Keating's view, the windows lie along articulatory dimensions which are supposed to be related to the phonological features. The mapping between features and articulatory dimensions is also part of Phonetic Encoding. Some ideas for relating phonological features to articulatory gestures have been proposed by Zsiga (1997). We speculate that this mapping itself can also be sensitive to context. Consider, as an example, the realization of [-voice] in terms of dimensions of glottal articulation, as discussed by Keating (1984, 1990b). For stops, this relation depends on prosodic structure: In some contexts, [-voice] is mapped to glottal spreading, in other contexts to glottal constriction; in other contexts to no articulation. [+voice] can be realized by a large set of articulatory parameters, including supra-glottal ones like pharyngeal expansion. The mappings also depend on the context provided by the other features in the same segment, e.g. while stops vary depending on prosody, voiceless fricatives alway have glottal spreading, because of the airflow requirement for fricatives.

Note that the view espoused here, that all aspects of Phonetic Encoding can be sensitive to prosodic context, takes much of the burden of phonetic realization off of so-called extrinsic allophones (which are defined by changes in symbolic features and so would be part of Phonological Encoding) and transfers it to the gradient phonetics. There seems to be no theoretical cost to phonological feature changes that produce extrinsic allophones; e.g. distinguishing [-back] and [+back] allophones of /l/, as is traditionally done, is natural in terms of features since the feature Back is already available in the phonology. However, assigning such variation instead to the gradient phonetics accords with our understanding of the phonetic facts (e.g. discussion of /l/ in section 4.4) and perceptual facts such as described by Manuel (1995) and Gow (in press). Keating (1996) notes that while categorical assimilatory allophones can be found, gradient phonetic assimilation seems to be much more common. An example given there is a distinction between various Russian vowel allophones, most of which seem to be gradient assimilations, but two of which (the [+back] allophone of /i/, and [-back] allophone of /a/) show categorical phonetic behavior.

6. Conclusion

L89/LRM99's approach to Word Form Encoding sees this process as the task of integrating segmental and prosodic information to determine the surface phonetic form of the words in an utterance. This approach has led to a number of important advances, in a framework that is shaped by three key design decisions: adoption of incremental encoding in PWd-sized units, separate (and later) generation of higher-level prosodic structures with subsequent integration of the two kinds of representation, and phonetic-encoding via retrieval of syllable-sized gestural scores from a precompiled syllabary. In combination with constraints derived from psycholinguistic experiments using e.g. priming, as well as results from computer implementation of the single-word aspect of the model, this line of inquiry has produced a number of important insights into the human Word Form Encoding process. For example, it provides a principled account of the time course of form encoding, which was strikingly missing from earlier accounts, and for some aspects of phonological variation in context, such as syllabification across certain kinds of word boundaries; and it allows for the expression of some aspects of phonetic variation in context, such as CV and VC coarticulation within the syllable. The model also describes an explicit mechanism for the serial ordering of already-ordered sublexical components, and for the delay in the onset of production pending form information about the entire PWd.

These design decisions, however, have left a number of unresolved problems. In general, these problems can be characterized in the following way: a number of speech phenomena necessitate small retreats from the principles of incrementality etc. adopted in the model; these departures, which rely on claims about lookahead/up/inside only in special cases, are not generous enough to account for the full range of effects; moreover, the details of how these departures work are often not fully specified. We believe that as these details are explored, the basic assumptions will be weakened as the advantages of strict incrementality etc. become less clear, and the model will begin to look more one that is consistent with the 'prosody first, segments later' approach we sketched in section 5. In this approach, looking ahead, up and inside are not infrequent, out-of-the-mainstream effects but are rather an integral part of the model, since segmental and prosodic processing mechanisms are deeply entwined.

Our sketch of a model based on these premises provides solutions for a number of problems posed by L89/LRM99's choices about how to model the process of integrating word form information with higher-level prosody. The availability of word-form information about upcoming words enables an account of cross-PWd sublexical errors, cross-PWd prosodic constraints and cross-PWd segmental context effects. The availability of higher-level prosodic structure makes possible an account of boundary- and prominence-governed phonological and phonetic effects. And integrating prosodic structure with phonetic encoding provides a more natural framework for computing the differential effects of prosody on different subconstituents within the PWd and syllable.

Two further points must be made. The first is that we still have a long way to go before we understand how speakers integrate prosodic and segmental information in Word Form Encoding. One striking example concerns the possible differences in this process among languages and even among dialects. LRM99 documents a number of interesting contrasts between languages that seem to suggest different processing, such as the more active role of syllables in French than in English and Dutch, and the preferential occurrence of segmental interaction errors between word onset consonants in some languages (e.g. English) but not in others (e.g. Spanish). Differences between dialects of a single language are also emerging. For example, for British English, LRM99 argue that an aspirated pronunciation of a voiceless final stop suggests resyllabification into the onset position both for a following vowel-initial inflection, in phrases such as *escorting*, and for a following vowel-initial clitics, as in *escort us*. In contrast, Hayes (1989) suggests that American English speakers can treat these two kinds of structures differently, lightly aspirating the /t/ before an inflection (e.g. in *visited*) but always flapping into the clitic (e.g. in *visit it*). He attributes this difference to contrasting prosodic structures for the two types of syntactic structures. Shattuck-Hufnagel (to appear) reports acoustic evidence to support this observation, at least for some speakers. Eventually, it will be important to model the ways in which the prosodic structures used by speakers of different languages and dialects may result in different effects on the surface phonetic forms of words, just as we need to address how different prosodic structures for different utterances of the same sentence in a single language may result in different surface phonetic word forms. This kind of cross-language processing distinction has not yet been addressed in detail by any model, so it is not yet possible to evaluate how well it can be dealt with by one approach vs. another. Like LRM99, we will leave this issue for future work.

Finally, it must be noted that the detailed LRM99 model of single word Phonological Encoding, along with its extensive integration with experimental data and its sophisticated handling of time course information, is the result of a mind-boggling amount of effort at the Nijmegen laboratory over the past decade to understand the complex workings of the human speech production system. The explicitness and comprehensiveness of this model have inspired, and continue to inspire, a generation of cognitive scientists and linguists to tease out and test its many implications, and to develop alternatives. Any proposal for a 'prosody first, segments later' model will have to account for the large body of relevant experimental data from this and other laboratories, from studies of language breakdown in aphasia and of language acquisition, as well as for the insights that have come from the development of formal phonological theories. The less-detailed L89 model of connected speech planning, in turn, tackles the even more challenging problem of understanding how speakers generate phrase-level prosodic structure, and how they specify the potentially extreme but highly systematic phonetic variation that has long been observed in continuous speech, particularly conversational speech. The work of the next decade will be to bring speech planning models to the admirable level of completeness exhibited in LRM99, without giving up the ambitious goals of L89. Such an accomplishment would have implications for many fields other than cognitive modeling of speech production, including automatic speech synthesis and, eventually, remediation of speech pathologies and the teaching of second languages. We look forward to these developments with eager expectation.

References

Baumann, M. (1995) The production of syllables in connected speech. Nijmegen U. dissertation.

Beckman, M. E., Swora, M. G., Rauschenberg, J. & de Jong, K. (1990) Stress shift, stress clash and polysyllabic shortening in a prosodically annotated discourse. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, 1: 5-8.

Beckman, M. & Pierrehumbert, J. (1986) Intonational structure in Japanese and English. *Phonology Yearbook 3*: 255-309.

Berkovits, R. (1993) Progressive utterance-final lengthening in syllables with final fricatives. *Language and Speech 36*: 89-98.

Bolinger, D. (1958) A theory of pitch accent in English, *Word 14*: 109-149.

Booij, G. (1985) Coordination reduction in complex words: a case for prosodic phonology. In H. van der Hulst and N. Smith (eds.), *Advances in Nonlinear Phonology*, Dordrecht: Foris, pp. 143-160.

Browman, C. P. & Goldstein, L. (1988) Some notes on syllable structure in articulatory phonology. *Phonetica 45*: 140-155

Browman, C. P.& Goldstein, L. (1990) Representation and reality: physical systems and phonological structure. *J Phonetics 18*: 411-424.

Browman, C. P.& Goldstein, L. (1992) Articulatory Phonology: An overview. *Phonetica 49*: 155-180.

Butterworth, B. (1989) Lexical access in speech production. In W. Marslen-Wilson (ed), *Lexical representation and processes*. Cambridge, Mass: MIT Press

D. Byrd & E. Saltzman (2002) The elastic phrase: Dynamics of boundary-adjacent lengthening. USC and Boston U. Ms.

Byrd, D., Kaun, A., Narayanan, S., and Saltzman, E. (2000) Phrasal signatures in articulation. In M. B. Broe and J. B. Pierrehumbert (eds.) *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, pp. 70-87. Cambridge University Press.

Cambier-Langeveld, T (2000) *Temporal Marking of Accents and Boundaries*. U. Amsterdam. LOT dissertation 32, *HIL 50*

Cho, T. (2001) *Effects of prosody on articulation in English*. UCLA dissertation.

Cohn, A. (1990) *Phonetic and Phonological Rules of Nasalization*. *UCLA Working Papers in Phonetics 76*

Crompton, A. (1982) Syllables and segments in speech production. In A. Cutler (ed), *Slips of the Tongue and Language Production*. Berlin: Mouton.

Crystal, D. (1969) *Prosodic systems and intonation in English*. London: Cambridge U. Press.

de Jong, K. (1995) The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation, *JASA 97*: 491-504

Dell, G. (1986) A spreading activation model of retrieval in language production. *Psychological Review 93*: 283-321.

Dell, G. (2000) Commentary: Counting, connectionism, and lexical representation. In M. B. Broe and J. B. Pierrehumbert (eds*.) Papers in Laboratory Phonology V: Acquisition and the Lexicon*, 335-348. Cambridge University Press.

Dell, G., Burger, L.K. and Svec, W.R. (1997) Language production and serial order: A functional analysis and a model. *Psychological Review 104*: 123-147

Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M. (1996) Glottalization of word-initial vowels as a function of prosodic structure. *J Phonetics 24*: 423-444.

Edwards, J., Beckman, M.E & Fletcher, J. (1991) The articulatory kinematics of final lengthening. *JASA 89*: 369-82.

Epstein, M. (2002) *Voice Quality and Prosody in English*. UCLA dissertation.

Ferreira, F. (1993) Creation of prosody during sentence production. *Psychological Review 100*: 233-253

Fougeron, C. (1998) *Variations Articulatoires en Début de Constituants Prosodiques de Différents Niveaux en Français*. U. Paris III – Sorbonne Nouvelle dissertation.

Fougeron, C. (1999) Prosodically conditioned articulatory variations: a review. *UCLA Working Papers in Phonetics 97*: 1-74.

Fromkin, V. A. (1971) The non-anomalous nature of anomalous utterances. *Language 47*: 27-52.

Fromkin, V.A. (1973) Introduction. In Fromkin, V.A. (ed), *Speech Errors as Linguistic Evidence,* 11-45. The Hague: Mouton.

Fry, D. (1969) The linguistic evidence of speech errors. *BRNO Studies of English 8*: 69-74.

Fujimura, O. (1993) C/D Model: a computational model of phonetic implementation, *Cognitive Science Technical Report* #5, Center for Cognitive Science, Ohio State University

Fujimura, O. (2000) The C/D model and prosodic control of articulatory behavior. *Phonetica 57*: 128-138.

Garrett, M.F. (1975) The analysis of sentence production. In Bower, G.H. (ed), *The Psychology of Learning and Motivation*, 133-177. NY: Academic Press

Garrett, M.F. (1976) Syntactic processes in sentence production. In Wales, R.J. and Walker, E. (eds) *New approaches to language mechanisms*, 231-256. Amsterdam: North Holland

Garrett, M.F. (1984) The organization of processing structure for language production: Applications to aphasic speech. In D. Caplan and A.R. Lecours (eds), *Biological perspectives on language*, 172-193. Cambridge, Mass: MIT Press.

Gee and Grosjean (1983) Performance Structures: A Psycholinguistic and Linguistic Appraisal. *Cognitive Psychology 15*: 411-458.

Gick, Bryan (1999) *The Articulatory Basis of Syllable Structure: A Study of English Glides and Liquids*. Yale U. dissertation.

Gow, D. (in press) Assimilation and anticipation in continuous spoken word recognition. *J Memory and Language*

Guenther, F.H. (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review 102*: pp. 594-621.

Hall, T.A. and Kleinhenz, U. (eds) (1997) *Studies on the Phonological Word*. Amsterdam and Philadelphia: John Benjamins.

Halliday, M. A. K. (1967) *Intonation and grammar in British English*. The Hague: Mouton.

Hayes, B. (1984) The phonology of rhythm in English. *Linguistic Inquiry* 15: 33-74

Hayes, B. (1989) The prosodic hierarchy in meter. In Kiparsky, P. and Youmans, G. (eds), Rhythm and Meter, Phonetics and Phonology 1, Academic Press.

Hayes, B. and Lahiri, A. (1991) Bengali intonational phonology. *Natural Language and Linguistic Theory 9*: 47-96.

Horne, M. (1990) Empirical evidence for a deletion formulation of the rhythm rule in English. *Linguistics 28:* 959-981.

Inkeles, S. and Zec, D.(eds) (1990) *The Phonology-Syntax Connection.* Chicago and London: U. of Chicago Press

Inkelas, S. And Zec D. (1995) Syntax-morphology interface. In Goldsmith, J. (ed.) *The Handbook of Phonology*. Oxford: Blackwell, pp. 535-49

Jun, S. (1993) *The phonetics and phonology of Korean prosody*. OSU dissertation.

Jurafsky, D., Bell, A., and Girand, C. (in press) The Role of the Lemma in Form Variation. To appear in Warner, N. and Gussenhoven, C. (eds.), *Papers in Laboratory Phonology VII*

Kaisse, E. (1985) *Connected speech: the interaction of syntax and phonology*. Academic Press.

Keating, P. (1984) Phonetic and phonological representation of stop consonant voicing. *Language 60*: 286-319

Keating, P. (1990a) The window model of coarticulation: articulatory evidence. In J. Kingston & M. Beckman (eds.) *Papers in Laboratory Phonology I* , Cambridge University Press, pp. 451-470.

Keating, P. (1990b) Phonetic representations in a generative grammar. *J. Phonetics 18*: 321-334

Keating, P. (1996) The Phonology-Phonetics Interface. In U. Kleinhenz (ed.) *Interfaces in Phonology, Studia grammatica 41*, Akademie Verlag, Berlin, pp. 262-278.

Keating, P., Cho, T., Fougeron, C. and Hsu, C. (in press) Domain-initial strengthening in four languages. To appear in *Papers in Laboratory Phonology VI*, Cambridge U. Press.

Kempen, G. & Hoenkamp, E. (1987) An incremental procedural grammar for sentence formulation. *Cognitive Science 11*: 201-258.

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

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991) The time course of lexical access in speech production: A study of picture naming. *Psychological Review 98*: 122-142.

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

Liberman, A. M., Cooper, F. S. Shankweiler, D. P., & M. Studdert-Kennedy (1967) Perception of the speech code. *Psychological Review 74*: 431-461.

Liberman, M. (1975) *The Intonational System of English*. MIT dissertation.

MacKay, D. G. (1972) The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology 3*: 210-227

Manuel, S. Y. (1995) Speakers nasalize /th/ after /n/, but listeners still hear /th/ . Journal of Phonetics 23: 453-476

McHugh, B. (1990) *Cyclicity in the phrasal phonology of Kivunjo Chaga*. UCLA dissertation.

Meijer, P. J. A. (1994) Phonological encoding: The role of suprasegmental structures. Nijmegen U. dissertation.

Meijer, P. J. A. (1996) Suprasegmental structures in phonological encoding: The CV structure. J. Mem. Lang. 35: 840-53.

Monaghan, A. I. C. (1990) Rhythm and stress-shift in speech synthesis. *Computer Speech and Language 4*: 71-78.

Nespor, M. and Vogel, I. (1986) *Prosodic phonology*. Foris.

Nespor, M. and Vogel, I. (1989) On clashes and lapses. *Phonology 6*: 69-116.

Peperkamp, S.A. (1997) *Prosodic Words*. Den Haag: Holland Academic Graphics – HIL dissertations 34

Pierrehumbert, J. B. (1980) *The Phonetics and Phonology of English Intonation*. MIT dissertation.

Pierrehumbert, J. (2001) Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. and P. Hopper (eds.), Frequency effects and the emergence of linguistic structure. John Benjamins, Amsterdam, 137-157

Pierrehumbert, J. (in press) Word-specific phonetics. In C. Gussenhoven and N. Warner (eds), *Papers in Laboratory Phonology VII*. Berlin: Mouton de Gruyter

Pierrehumbert, P. and Talkin, D. (1992) Lenition of /h/ and glottal stop. In G. Docherty and D.R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, pp. 90-117. Cambridge U. Press.

Plaut, D. C. & Kello, C. T. (1998) The emergence of Phonology from the Interplay of Speech Comprehension and Production: A Distributed Connectionist Approach. In B. MacWhinney (ed.) *The emergence of language*. Mahweh, NJ: Erlbaum.

Price, P., M. Ostendorf, S. Shattuck-Hufnagel and C. Fong (1991) The use of prosody in syntactic disambiguation. *J. Acoust. Soc. Am. 90*: 2956-2970.

Redi, L. and S. Shattuck-Hufnagel (2001) Variation in the realization of glottalization in normal speakers. *J. Phonetics 29*: 407-429

Roelofs, A. (1992) *Lemma retrieval in speaking: A theory, computer simulations, and empirical data*. U. Nijmegen dissertation.

Roelofs, A. (1997) The WEAVER model of word-form encoding in speech production. *Cognition 64*: 249-84.

Schiller, N. (1997) *The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography*. Nijmegen U. dissertation.

Schiller, N. (1998) The effect of visually masked syllable primes on the naming latencies of words and pictures. *J. Mem. Lang. 39*: 484-507.

Selkirk, E. (1984) *Phonology and syntax: the relation between sound and structure*. MIT Press.

Selkirk, E. (1986) On derived domains in sentence phonology. *Phonology Yearbook 3*: 371-405.

Selkirk, E. (1995) Sentence prosody: intonation, stress and phrasing. In Goldsmith, J.A. (ed), *The Handbook of Phonological Theory,* 550-569. Cambridge, Mass: Blackwell.

Sevald, C. A., Dell, G., and Cole, J. (1995) Syllable structure in speech production: Are syllables chunks or schemas? *J Mem Lang 34*: 807-820

Shattuck-Hufnagel, S. (1979) Speech errors as evidence for a serial-ordering mechanism in sentence production. In: *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*, W. E. Cooper & E, C. T. Walker (eds.). Lawrence Erlbaum.

Shattuck-Hufnagel, S. (1992) The role of word structure in segmental serial ordering. *Cognition 42*: 213-259.

Shattuck-Hufnagel, S. (forthcoming) A prosodic view of phonological and phonetic encoding. To appear in proceedings of LabPhon8 (June 2002).

Shattuck-Hufnagel, S. and Turk, A. (1996) A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguistic Psychology 25*: 193-247.

Shattuck-Hufnagel, S., M. Ostendorf and K. Ross (1994) Stress shift and early pitch accent placement in lexical items in American English. *J. Phonetics 22*: 357-388

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992) TOBI: A Standard for Labeling English Prosody. In *Proceedings of the 1992 ICSLP*, vol. 2, pp. 867-870.

Sproat, R. and Fujimura, O. (1993) Allophonic variation in English /l/ and its implications for phonetic implementation. *J Phonetics 21*: 291-311

Stemberger, J. P. (1985) An interactive activation model of language production. In A.W. Ellis (ed), *Progress in the psychology of Language I*, Hillsdale, NJ: Erlbaum

Stemberger, J.P. (1991) Radical underspecification in language production. *Phonology 8*, 73-112

Steriade , D. (2000) Paradigm Uniformity and the Phonetics/Phonology Boundary. In J. Pierrehumbert and M.Broe (eds.) *Papers in Laboratory Phonology V*, Cambridge U.

Turk, A. (1999) Structural influences on boundary-related lengthening in English. *Proceedings of 14th ICPhS*, vol. 1, p. 237-240.

van Turennout, M. Hagood, P., and Brown, C. M. (1997) Electrophysiological evidence on the time course of semantic and phonological processes in speech production. *J. Exper. Psych.: Learning, Memeory and Cognition 23*: 787-806.

van Turennout, M. Hagood, P., and Brown, C. M. (1998) Brain activity during speaking: From syntax to phonology in 40 msec. *Science 280*: 572-74

Vousden, J.I., Brown, G.D.A. and Harley, T.A. (2000) Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology 41*: 101-175

Watson, D. (2002) Intonational phrasing in language production and comprehension. Forthcoming MIT dissertation.

Wheeldon, L.(2000) Generating prosodic structure. In Wheeldon, L. (ed), *Aspects of Language Production*, 249-274. Hove, East Sussex: Psychology Press.

Wheeldon, L. R. & Lahiri, A. (1997) Prosodic units in speech production. *J Memory and Language 37*: 356-381.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (1992) Segmental durations in the vicinity of prosodic phrase boundaries. *JASA 91*: 1707-17.

Wright, R. (in press) Factors of lexical competition in vowel articulation. To appear in *Papers in Laboratory Phonology VI*, Cambridge U. Press.

Zsiga, L. (1997) Features, gestures, and Igbo vowel assimilation: An approach to the phonology/phonetics mapping. *Language 73*: 227-74

Zue, V. and S. Shattuck-Hufnagel (1979) The palatalization of voiceless alveolar fricative /s/ in American English. In Fischer-Jorgensen (ed.), *Proc. IXth International Congress of Phonetic Sciences*, Vol. 1, p. 215