Prosodic Constraints and Processing Theory
Author(s): Lori Taft
Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society (1985), pp. 519-528

Prosodic Constraints and Processing Theory

Lori Taft
University of Pittsburgh

1. Introduction.

One of the key questions facing psycholinguists involved in language processing concerns the nature of interaction between theories of processing and theories of grammar.

There are extreme positions on this question. At one extreme is the view that there is absolutely no relationship between the structures which offer the best linguistic description of a language and the products of language processing. In this view, two grammars are needed: a grammar of competence, which represents a language user's knowledge, and a grammar of performance, which characterizes the structures and constraints relevant to processing.

At the other extreme is the position that there is a one-to-one correspondence between, on the one hand, the structures and constraints which best characterize language and, on the other, the processes and representations used in language understanding.

A position between these two extremes acknowledges the relationship between a theory of grammar and theory of processing, and seeks to find that relationship. It holds that there is no simple one-to-one mapping between linguistic structures and constraints, and psychological structures and processes, but that a mapping does exist.

This is the position I take in this paper. With the basic assumption that the structures and constraints proposed within linguistic theory are relevant to processing theory, the task before us is to specify how they are relevant.

1.1. The role of phonological constraints.

My focus here is on phonological constraints and their possible role in the processing of spoken language. In particular, I am concerned with the use of phonological knowledge at the stage of lexical access: that is, contacting stored representations of lexical material.

It may be argued that phonological knowledge is not necessary for making decisions about lexical items before their representations are contacted. Since the set of items of a language is finite, it might in principle be possible to store all conceivable acoustic representations of a word, and simply match the incoming acoustic material to stored representations.

However, there is so much variability in the acoustic signal, due for

example to speaker differences, speech rate differences, and the syntactic context of a word, that this appears to be an inadequate characterization of how we process lexical material. In addition, the acoustic variants of a lexical item do not comprise an arbitrary collection of phonetic forms. Rather, they are tokens which are related to each other and to their underlying form in systematic ways. Part of knowing a language is knowing these systematicities in how phonetic forms are related to their underlying forms.

Given these considerations, it is plausible that the hearer calls upon that knowledge to impose an interpretation on incoming speech. That is, just as the listener must call upon syntactic knowledge to interpret incoming material above the word level, he may call upon phonological knowledge to interpret incoming material at or below the word level.

With that established, we can formulate two broad questions. First, at what level of processing are phonological constraints used? Second, what exactly are the phonological constraints which are relevant to processing? As an answer to the first question, I'd like to consider a place where constraints on lexical stress may play a role in processing. In particular, I'd like to focus on the process of word boundary identification, or what I will call "lexical parsing". As an answer to the second question, I'd like to consider theories of suprasegmental phonology, and in particular, theories of metrical structure as presented in Selkirk (1980) and Hayes (1980).

1.2. The segmentation problem.

To identify words, the listener must segment the continuous acoustic signal into discrete units. This would be trivial if speakers regularly left silences between words; the speaker, however, is rarely so accomodating to the listener. Further, if every language provided reliable and consistent phonetic cues to word boundaries, lexical segmentation would not pose a problem for listeners. However, though cues to word junctures are found in some languages, none are systematically present in all languages, nor (to my knowledge) systematically present in all utterances within one language.

The segmentation problem is complicated by the temporal nature of the incoming signal. The hearer does not receive information about all parts of an utterance simultaneously. If that were the case, it would be possible to exhaustively parse the signal by imposing an analysis which would conform to the constraints of the language. Instead, the listener must assign meaning to portions of the signal as they are received.

It has been shown (Marslen-Wilson and Welsh, 1978; Cole and Jakimik, 1978) that listeners do not wait until the entire signal corresponding to a word has been heard before interpreting the signal. Thus, efficient lexical access

depends on accurately identifying word onsets and finding the correct candidate from a set of plausible alternatives. Considerable experimental evidence supports the contention that the identification of word onsets plays a distinguished role in word identification (Cole, 1973; Cole and Jakimik, 1980; Marslen-Wilson, 1975; Marslen-Wilson and Welsh, 1978). However, though those studies demonstrate the importance of word onsets, they do not directly address the question of how onsets are identified.

2. Prosody in lexical parsing.

Prosody provides a potentially rich source of information to the listener. Stressed syllables are more salient than unstressed syllables and provide better acoustic information than unstressed syllables. In fact, there is evidence that listeners anticipate the arrival of stressed syllables (Cutler, 1976) and may structure the signal into rhythmic units which either begin or end with a stressed syllable (Martin, 1972). Thus, prosody is potentially useful for structuring the input and guiding the listener to the more important parts of the signal.

However, the role of prosody in onset identification has not been systematically explored. This follows in part from assumptions about the representation of lexical items. Current lexical access models assume that words are represented as concatentations of segments, with no structure internal to words (except possibly morphological structure). In none of the models is an explicit claim made about phonological representations. The implicit assumption appears to be that those representations are linear sequences of phonemes, represented as feature bundles, approximating the phonetic representations assumed in standard generative phonology (eg., Chomsky and Halle, 1968). Yet nothing in these theories forces a strictly linear view of word structure. Additional structure in lexical representations would simply be superfluous to current versions of these models, since at present they contain no parsing mechanisms to exploit the information about structure.

Given the consideration that prosody may play a role in access, we may now consider how the perceptual system might use prosodic constraints, as characterized first in the framework of Chomsky and Halle (1968), and then in a metrical phonology framework (e.g., Selkirk, 1980; Hayes, 1980).

2.1. Lexical access within alternative phonological frameworks.

Within the framework of SPE, lexical access can be characterized as moving backwards through a derivation, undoing each step to arrive at an underlying representation. Viewing lexical access in this way presents two problems.

One problem concerns the computational load that would result from undoing certain phonological rules. A case in point is the Main Stress Rule in English, a simplified version of which is given in (1).(Chomsky and Halle, 1968:77)

(1) Main Stress Rule

$$V \text{ ---> } [1\text{stress}] \,/\, \text{——} \; C_0 \left(\begin{bmatrix} \text{-tense} \\ V \end{bmatrix} C_0 \right)$$

$$/ \text{——} \; \langle \begin{bmatrix} \text{-tense} \\ V \end{bmatrix} C_0 \rangle \quad \langle N \rangle$$

The computation which would be required to undo this rule may be enormous. Since real-time constraints such as memory limitations play a role in lexical access, such considerations would certainly weaken the appeal of certain types of phonological rules such as the Main Stress Rule.

Another problem with the notion of undoing a derivation concerns the length and nature of the material, or the "window", which would be necessary for the listener in order to undo rules. Phonological rules operate on entire words (though perhaps affecting only one segment for any particular rule). The successive lines of a derivation contain the representation of the entire word, up to that point in the derivation. Thus, in undoing a derivation, the listener would need access to the entire word to undo each phonological rule. Evidence from psycholinguistic studies (Marslen-Wilson and Tyler, 1980) indicates that a word can often be recognized before all of it is heard. Thus, it cannot be that the access process awaits the phonetic representation of an entire word and only then starts to perform a reverse derivation.

Metrical theory as proposed in Selkirk (1979, 1980) and Hayes (1980) brought a change in the level of representation at which stress is specified. In

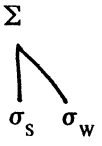
particular, Selkirk argued that the prosodic categories of syllable (σ), foot (Σ), and word (ω) are linguistically significant levels of phonological representation. It was argued that certain phonological processes have as their domain the syllable, foot, or word, and thus that the description of certain processes crucially must refer to these categories.

For English it was proposed that words be exhaustively parsed into "stress feet" in underlying representations, with the possible foot types given in (2).

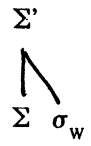
(2)

a. monosyllabic

b. bisyllabic



c. superfoot



If syllables and feet constitute the domain for certain phonological processes, then information about these units may be extractable from the incoming signal, rather than being derived from an underlying representation. It is thus possible that the listener is constructing syllables and feet and submitting those units to the lexicon to compare with stored representations. Since words are assumed to be stored with information about foot structure, this provides a way of matching the incoming signal with stored representations at a level intermediate between the segment and the word.

In the next section, I consider the hypothesis that the listener is indeed structuring the incoming signal into rhythmic units and that this structuring is providing initial guesses about where word boundaries are located in the incoming signal.

3. Prosody-based lexical parsing strategies.

When an utterance first reaches a hearer, the onset of the first word can be easily postulated - it will be identical to the onset of the entire utterance. This onset can be submitted to the lexicon for comparison with stored lexical representations. It is not clear just how much information about a word is available at this point, but I assume that at the least, a phonological representation is accessed. This is compared to the incoming signal. Whatever information is included in the phonological representation, it is potentially usable in the mismatch procedure. In particular, I hypothesize that suprasegmental structure is available.

If feet are being constructed, then they provide a unit of matching with stored representations. That is, the "window" used for matching the incoming signal to stored representations is defined structurally (i.e., in terms of syllables and feet) rather than strictly temporally (i.e., in terms of a window of some specified amount of time). It will be the onsets of feet, i.e., strong syllables, which are the most salient portions of the signal. A salience-based strategy for onset identification could thus be formulated as in (3).

(3) *Salience to Onset Strategy (SOS)*

Use the salient portions of the incoming signal, plus the prosodic constraints of the language, to find word onsets.

For English, the SOS will take stressed syllables as the salient portions of the signal. Given the constraints on foot structure in English, stressed syllables are taken to be word onsets. The strategy for English is given in (4).

(4) *Salience to Onset Strategy (English)*

Hypothesize a word onset at each stressed syllable.

I have hypothesized that prosodic structures are imposed on the incoming signal. These structures are consistent with the language-specific constraints on prosodic structures. The formulation of this strategy is given in (5).

(5) *Prosodic Domain Strategy (PDS)*

Segment the incoming signal into prosodic units which are well-formed according to the constraints of the language. Submit these units to the lexicon for comparison with stored lexical material.

The language-specific version of the PDS defines the prosodic domain which is relevant to lexical access for that language. For English, I claim that the relevant prosodic domain is the stress foot. This means that the listener imposes a foot structure on the signal before making decision about lexical segmentation. The strategy for English is given in (6).

(6) *Prosodic Domain Strategy (English)*

Segment the incoming signal into feet. Submit these units to the lexicon for comparison with stored lexical material.

Elsewhere (Taft, 1984) I have presented experimental evidence supporting the SOS(E). One experiment tested preferred segmentations of phonetically ambiguous items (e.g., *lettuce - let us*; *incite - in sight*). The results were consistent with the hypothesis that listeners take strong syllables to be word-

initial, and do *not* take weak syllables to be word-initial unless they are clearly marked (e.g., are utterance-initial). Another experiment tested lexical access times for bisyllabic words whose stress pattern was pronounced either correctly (e.g., *CACtus, susPENSE*) or incorrectly (e.g., *cacTUS, SUSpense*). The results showed differential effects of a stress mispronunciation depending on whether it moved stress toward or away from the word onset. Thus, *CACtus* showed faster access times than *cacTUS*, but *SUSpense* showed faster access times than *susPENSE*. (However, only the first type of mispronunciation resulted in a significant difference in access time.)

Evidence for the PDS is more tentative, and comes from two studies described in detail elsewhere (Taft, 1984). The PDS predicts that the prosodic structure of an utterance should make a difference in how it is initially segmented into words.

In particular for English, since a rhythmic foot boundary will not necessarily coincide with a word boundary, the PDS may lead to a wrong segmentation in cases where the foot boundary crosses a word boundary. Consider, for instance, the sequences of strong and weak syllables shown in (7) (with "#" indicating true word boundaries).

(7)
a. S W # W S
b. S W # S W
c. S W W # W S
d. S W W # S W

The PDS predicts that (a) will be more difficult to parse than (b). This is because the listener is constructing feet, and will miss the word-initial *weak* syllable on the second word of (a), but will not miss the word-initial *strong* syllable on the second word of (b). (By hypothesis, the principle is to construct maximal feet, i.e., superfeet, where possible.) In contrast, the PDS predicts that (c) should be no more difficult to parse than (d). In (c), the maximal foot coincides with the word boundary, so the word-initial weak syllable will be tried as a word onset. In (d), the initial syllable of the second word will be tried as an onset, because it is strong, and also because the preceding maximal foot coincides with the word boundary. Thus, there should be a greater difference in segmentation time for (a) *vs.* (b) than for (c) *vs.* (d).

In an experiment testing these predictions, I presented 4- and 5-syllable sequences of either "word+word" (W+W) or "nonword+word" (NW+W) to subjects. Their task was to decide whether they had heard a word+word or a nonword+word sequence and press the appropriate button on a panel in front of them. Response times were measured, and are reported in Table 1

for the four conditions in (7).

Table 3-1: Mean Reaction Times to Lexical Sequence Decision (msec.)

Rhythmic Sequence:	Predicted Differences:	W+W	NW+W
a. SW # WS	(a-b)>0	998	1177
b. SW # SW		947	1126
c. SWW # WS	(a-b)-(c-d)>0	946	1028
d. SWW # SW		932	987

Table 2 presents the differences across conditions highlighted in the above discussion.

Table 3-2: Differences across Conditions (msec.)

Differences:	W+W	NW+W
(a-b)	51	51
(c-d)	14	41
(a-b)-(c-d)	37	10

Interaction between the form of the first item and the form of the second item was not significant, but the pattern of results is striking. For all conditions, the reaction time differences were in the right direction, i.e., they patterned exactly as predicted by the PDS. This would not be expected if rhythmic structure played no part in lexical segmentation. Thus, though the results did not present striking confirmation of the PDS, they suggest that rhythm is indeed relevant to segmentation.

4. Implications for processing theory.

This work affects lexical access models in three ways. The first concerns the structure of lexical representations. I have argued that the "flat" segmental word structures assumed by auditory lexical access models are oversimplified

in that they ignore word-internal structure. Of course, the phonological theories make no claims about the "psychological reality" of the constructs they employ to describe suprasegmental phenomena. However, the positing of linguistically significant units such as syllables and feet is motivated partly by the fact that phonological processes operate within or across those domains. For the listener, this means that phonetic information about those prosodic units may be extracted from the signal, by considering whether a particular phonological process has or has not occurred.

In addition, the evidence presented here suggest that the process of phonologically structuring the input may represent a distinct level of processing. If the conclusions I have drawn are correct, then they force the models to take account of both phonological structure and the process by which word onsets are identified. In addition, evidence from Japanese, which I have reported elsewhere (Taft, 1984) shows different parsing preferences from speakers of dialects which differ in permissible tone melodies on words. These experimental results demand the conclusion that phonological knowledge is important in processing.

Finally, theories of suprasegmental phonology provide a framework for formulating questions concerning the nature of the prosodic constraints relevant to processing, and how those constraints are used in understanding fluent speech.

Bibliography

Chomsky, N. and M. Halle. *The Sound Pattern of English.* : Harper and Row 1968.

Cole, R. Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, 1973, *13*, 153-156.

Cole, R. and J. Jakimik. Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of Information Processing,* : Academic Press, 1978.

Cole, R. and J. Jakimik. How are syllables used to recognize words? *Journal of the Acoustical Society of America*, 1980, *67*, 965-970.

Cutler, A. Phoneme monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, 1976, *20*, 55-60.

Goldsmith, J. An overview of autosegmental phonology. *Linguistic Analysis*, 1976, *2*, 23-68.

Hayes, B. A Metrical Theory of Stress Rules. Published by Indiana University Linguistics Club, 1980.

Liberman, M. and A. Prince. On stress and linguistic rhythm. *Linguistic Inquiry*, 1977, *11*, 511-562.

Marslen-Wilson, W. Sentence perception as an interactive parallel process. *Science*, 1975, *189*, 226-228.

Marslen-Wilson, W. and A. Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 1978, *10*, 29-63.

Marslen-Wilson, W. and L. Tyler. The temporal structure of spoken language understanding. *Cognition*, 1980, *8*, 1-71.

Martin, J. Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, 1972, *79*, 487-509.

Mehler, J. The role of syllables in processing: infant and adult data. *Philosophical Transactions of the Royal Society of London*, 1981, *3295*, 119-138.

Mehler, J., Y. Y. Domergues, U. Frauenfelder, and J. Segui. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 1981, *20*, 298-305.

Selkirk, E. The role of prosodic categories in English word stress. *Linguistic Inquiry*, 1980, *14*, 563-605.

Selkirk, E. On prosodic structure and its relation to syntactic structure. Paper presented at the Conference on the Mental Representation of Phonology, University of Massachusetts, Amherst, 1979. Distributed by Indiana University Linguistics Club, 1980.

Studdert-Kennedy, M. Speech Perception. In N. Lass (Ed.), *Contemporary Issues in Experimental Phonetics*, : Academic Press, 1976.

Taft, L. *Prosodic Constraints and Lexical Parsing Strategies.* PhD thesis, University of Massachusetts at Amherst, 1984.