Thoughts on Prosodic Structure

Daniel J. Hirst

Thoughts on Prosodic Structure*

Hirst, Daniel

Abstract: Our ideas about prosodic representation are heavily influenced by our knowledge of written language. All writing systems represent utterances as a linear sequence of elements drawn from a finite set of characters. In many languages, special characters such as spaces or punctuation marks are used as boundary symbols. There is a general consensus today that utterances, although themselves produced, transmitted and perceived as a linear stream of (respectively) physiological, acoustic and perceptual events, are mentally represented as a prosodic structure in which smaller chunks of speech are grouped into larger chunks following a hierarchy of phonological levels, and that this hierarchy is only partially related to the more abstract syntactic structure. In this paper, I present and discuss some thoughts on the nature of these prosodic chunks and the ways in which prosodic structure differs both from written language and syntactic structure. I suggest, in particular, that a less linear approach to prosodic structure may lead to significant and sometimes surprising insights into the nature of prosodic representations.

Keywords: prosodic structure; phone; mora; syllable; rhythm; melody; intonation

1. Introduction

There is a general consensus today that, although utterances are produced, transmitted and perceived as a linear stream of (respectively) physiological, acoustic and perceptual events, see Figure (1), they are mentally represented as a hierarchical

* This paper is based on the keynote presentation I gave at the 9th International Conference on Speech Prosody in Poznań Poland, June 13, 2018. A version of the presentation was also given as an ISCA distinguished lecture at Fudan University, Shanghai on May 23, 2019.

prosodic structure, in which smaller chunks of speech are grouped into larger chunks following a hierarchy of phonological constituents, and that this prosodic structure is only partially related to the more abstract syntactic structure.

A classic view of prosodic structure in English, for example, which can be traced back to work by British phoneticians in the early twentieth century, is that *utterances* are made up of groups of *intonational phrases*, intonational phrases are made up of groups of *stress feet*, stress feet are made up of groups of *syllables* and syllables themselves are made up of groups of *phones*, resulting in a hierarchical structure like that in Figure (2).

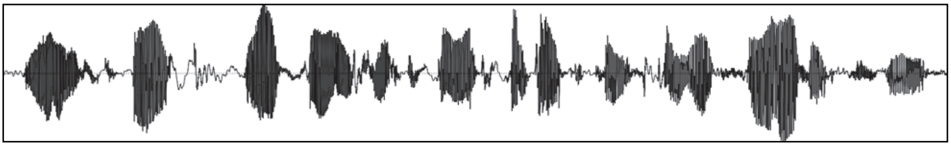
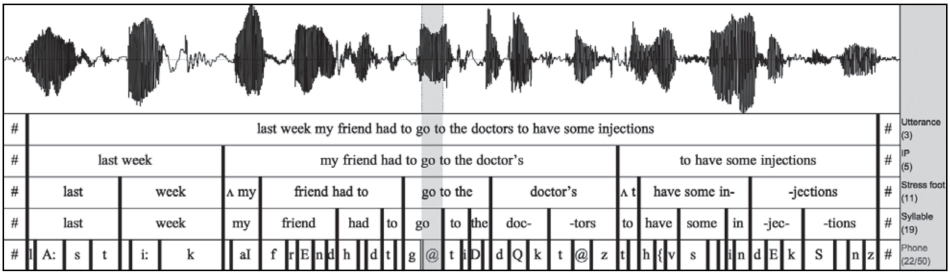


Figure 1 The acoustic signal (waveform) for the utterance "*Last week, my friend had to go to the doctor's, to have some injections*".



# last week my friend had to go to the doctors to have some injections # Utterance (3)
# last week my friend had to go to the doctor's to have some injections # IP (5)
# last week ʌ my friend had to go to the doctor's ʌ t have some in- -jections # Stress foot (11)
# last week my friend had to go to the doc- -tors to have some in -jec- -tions # Syllable (19)
# l A: s t i: k aI f r E n d h d t g @ t i D d Q k t @ z t h { v s i n d E k S n z # Phone (22/50)

Figure 2 The acoustic signal (waveform) and prosodic structure for the utterance "*Last week, my friend had to go to the doctor's, to have some injections*", assuming the prosodic constituents: *phone*, *syllable*, *stress foot*, *intonational phrase* and *utterance*.

Ladd (2014) notes that work in the area of metrical phonology:

> …has led to a variety of theoretical ideas about constituent structure in phonology (...) whose potential has, in my opinion, only begun to be explored. [p. 50]

In the rest of this paper, I present and discuss some thoughts about the nature of phonological constituent structure.

2. Phones

The influence of our writing system, especially with alphabetic writing, is particularly noticeable in the representation of individual speech segments, or phones. In the English word "scrambling", for example, there is a nearly one-to-one correspondence between the 10 letters of the orthographic transcription of the word and the 9 symbols of the phonemic transcription /'skramblɪŋ/, the only exception being the final consonant of the word, /ŋ/, which corresponds to two letters "ng".

2.1 Consonant clusters

English word onset clusters can contain up to three consonants. There are, however, a large number of constraints as summarised in the following finite state diagram[1]:

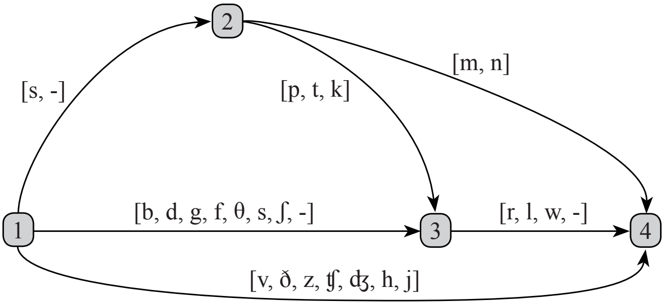


Figure 3 Finite state diagram of English onset consonant clusters

When there are three consonants, the first is necessarily /s/, the second a voiceless stop /p/, /t/ or /k/, and the third a sonorant /l/, /r/ or /w/.

If this were the whole story, we would expect to find 67 possible onset clusters (2*2 + 2*3*4 + 8*4 + 7).

In addition, however, there are other constraints not shown in the finite state diagram. If the stop is /p/, the sonorant cannot be /w/; if it is /t/, the sonorant cannot be /l/. In other words, (not counting /s/), there can be only one coronal [+cor] consonant in the cluster, and there can be only one labial consonant [+lab].

[1] Simplified from Coleman (2005:117). The symbol [-] indicates the possibility of an empty transition.

Two consonant onsets are either:

/s/ + {p, t, k, m, n}

or {p, t, k, b, d ,g, f, θ , s, ʃ} + {l, r, w}

Once again, labial stops or fricatives cannot be followed by /w/ and coronal stops and fricatives (except /s/) cannot be followed by /l/.

I suggested several years ago (Hirst, 1985) that the constraints on word onset clusters in English are so great that the clusters could in fact be analysed phonologically as underlying single segments, defined by a single array of distinctive features.

Since there are practically no restrictions on the consonants which can be preceded by /s/ or followed by /r/ (except for the impossible combination of the two /sr/), I took the additional step of suggesting that, in a cluster, /s/ and /r/ are unspecified for place features and represent respectively stridency [+str] and sonorance [+son].[1] This allowed me to define consonant clusters as complex segments, defined by a single set of features: [± cont; ± son; ± str; ± nas; ± voice, ± lab; ± cor; ± high].

It is not possible for a non-nasal segment to be at the same time a stop and a sonorant, so a feature set such as: [-cont +son -nas -voice +lab] has to be linearised as a voiceless labial stop followed by an unspecified sonorant /r/, which we can represent as /p^R/, which captures the idea that this is a labial stop with a sonorant release. For the cluster /pl/, we can simply add the [+cor] specification for the coronal sonorant, so that [-cont +son -voice +cor +lab] is linearised as /p^L/.

Similarly, since it is not possible for a consonant to be simultaneously [-cont] and [+str], a strident stop, the set [-cont +str +lab] is linearised as /${}^{S}p$/.

This leads to an inventory of 57 onsets for English (or 58 if we include the null onset), each defined by a unique single column of distinctive features, and which we can represent as:

(1) *a.* p, p^R, p^L, ${}^{S}p$, ${}^{S}p^R$, ${}^{S}p^L$, t, t^R, t^W, ${}^{S}t$, ${}^{S}t^R$, ${}^{S}t^W$, ʧ, k, k^R, k^L, k^W, ${}^{S}k$, ${}^{S}k^R$, ${}^{S}k^L$, ${}^{S}k^W$

b. b, b^R, b^L, d, d^R, d^W, ʤ, g, g^R, g^L, g^W

c. f, f^R, f^L, θ, θ^R, θ^W, s, s^L, s^W, ʃ, ʃ^R, ʃ^L

[1] When /s/ and /r/ occur by themselves as an onset, they obviously need to be specified for place of articulation, perhaps respectively [+cor] and [+cor+high].

d. v, v^R, v^L, ð

e. m, ^{S}m, n, ^{S}n

f. r, l, w, j, h

An inventory of 57 onsets is well within the range of the number of consonants classically described for different languages which, according to Maddieson (1984:7), ranges from 11 (Rotokas; Papua, New Guinea) to 141 (!Xóõ, Southern Khoisan; Botswana). In many of the languages with large consonant inventories, these include what Sagey (1986) has called *contour segments*, such as affricates or prenasalised stops. My analysis of English onset clusters is in effect a proposal to treat them all as contour segments.

The word "scrambling", then, at this point, could be represented as a sequence of 6 segments: /^{S}k^R a m b^L ɪ ŋ/

2.2 Velar nasals and nasal vowels

A feature like [+nasal] for vowels is obviously necessary to describe lexical distinctions in languages like French, Polish and Portuguese, which have lexically distinctive nasal vowels. In midi (=Southern) French, nasal vowels are often realised with a fully or partially de-nasalised vocalic portion followed by a nasal sonorant which is homorganic to the following consonant (if any). Thus in some varieties of this accent, instead of standard French:

(2) a. *camper* [kɑ̃pe], *chanter* [ʃɑ̃te], *planquer* [plɑ̃ke],

we can hear:

(2) b. *camper* [kampe], *chanter* [ʃante], *planquer* [plaŋke].

When there is no following consonant, the nasal vowel is often linearised as an oral vowel followed by a velar nasal, so that instead of standard French:

(3) a. *banc* [bɑ̃], *bon* [bɔ̃], *bien* [bjɛ̃],

we can hear:

(3) b. *banc* [baŋ], *bon* [bɔŋ], *bien* [bjɛŋ].

Treating the final nasal velar as the linearisation of an underlying nasal vowel in midi French, provides a natural explanation for the otherwise unexplained fact that, in this language, the nasal velar cannot occur in syllable initial position.

The English velar nasal consonant /ŋ/ also has a defective distribution: it can only occur in "syllable-final" position, and never word (or syllable) initially[1]. To account for this, we could analyse the English velar nasal as the result of the linearisation of a feature [+nasal] on the vowel of the representation. In this account, a word like "sing" would be analysed as phonologically a CV sequence with a [+nasal] feature on the V segment. We can represent this as /sɪ^N/[2]. The co-occurrence constraints on nasals in intersyllabic position can then be easily accounted for as an underlying nasal vowel followed by a non-nasal consonant so that we would have *singer* ['sɪŋə] = /sɪ^{N}ə/, *hinder* ['hɪndə] = /hɪ^{N}də/, *timber* ['tɪmbə] = /tɪ^{N}bə/, *finger* ['fɪŋgə] /fɪ^{N}gə/.

The word "scrambling", could then be represented as a sequence of only 4 segments: /^{S}k^R a^N b^L ɪ^N/

3. Syllables

In most alphabetic writing systems, words are separated by spaces or punctuation marks. At the level of the syllable, there is not a necessary agreement between the division into syllables and the division into words, which are the units that are presumably stored in a "mental lexicon" and that need to be accessed in order to interpret the meaning of an utterance.

If we take the French sentence:

(4) a. *Il est en or.*

b. /i.lɛ.tɑ̃. nɔʁ/

It is in gold. = "It's made of gold."

Like (4a), which contains four words, (4b) contains four syllables. Not one of the

[1] This is a language specific characteristic — in other languages (such as Cantonese), /ŋ/ may occur as a syllable onset.

[2] The notation /ɪ^N/ is simply intended as a representation of underlying nasality and is perfectly equivalent to the notation / ɪ̃/. Here I use the superscripted capital N to keep the same type of notation I use for the complex consonants.

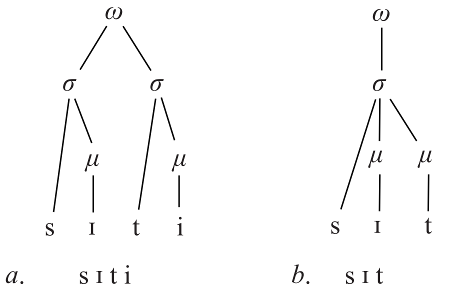
four syllables, however, corresponds exactly to one of the four words in the orthographic representation. The reason for this is that French makes regular use of *liaison* and *enchâinement* (linking), so that there is a strong tendency to "resyllabify" words whenever possible, to favor open syllable structure.

3.1 Moras

Between the phone and the syllable, many recent phonological analyses posit the existence of a *mora* which is essentially a unit of timing (cf. Hyman (1985)'s *weight units* (WUs), which he explicitly compares to the traditional concept of mora). .

Moraic Phonology (Hayes 1989; Broselow, 1996) defines representations, such as (5) for the English words "city" and "sit":

(5)



In which vowels and syllable-final consonants are linked to moras, whereas syllable-initial consonants are not linked to a mora (μ) but directly to a syllable node (σ).

The fact that syllable-initial consonants do not affect the weight of syllables is well known from studies of lexical stress systems, but as Allen (1973) noted:

> This approach, (...) leaves unexplained why even a single consonantal "surplus" following the vowel should create "length" of syllables just as an additional vowel mora, whereas any amount of consonantal surplus preceding the vowel (such as στρόφος) should be irrelevant. [p. 59]

A possible explanation for this asymmetry can be found in work on co-articulation. Öhman (1966), in a study of nonsense VCV words pronounced by English, Russian and Swedish speakers, found that there was a definite influence of the second vowel both on the intermediate consonant and on the initial vowel. He noted:

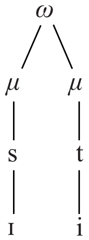
> We have clear evidence that the stop-consonant gestures are actually superimposed on a context-dependent vowel substrate that is present during all of the consonantal gesture. [p. 165]

This, together with many other studies, suggests that rather than being linearly ordered, as in the phonetic transcription of an utterance, the initial consonant(s) and the vowel in a short syllable are actually produced at the same time, but that the vocalic gesture continues after the end of the consonant.

This can easily be demonstrated by preparing to pronounce the words “sea” and “Sue”, without actually producing any sound. It will be noticed that the lips are spread for the first syllable but rounded for the second, even before any sound is produced. This shows clearly that the articulation of the vowel and that of the consonant must be prepared at the same time, before the articulation of either has begun.

In the framework I am outlining here, this simultaneous articulation is taken as central to the representation and I will assume that the onset consonants are in fact linked to the same mora as the “following” vowel. The mora is taken to represent an explicit timing slot, so that instead of (5a), we have something we can represent as:

(6)

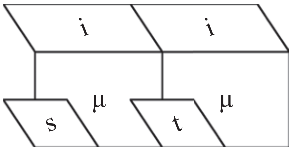


which is intended to represent the fact that both the vowel and the consonant are linked directly to the same mora.

In fact, with a representation like this, it is not clear that the syllable is necessary as a distinct level of phonological constituent.

We should think of consonants and vowels as occupying distinct tiers — so when I refer to a C segment or a V segment, this can be thought of as a set of distinctive features, defining a consonantal or a vocalic segment as discussed in the preceding section, possibly a contour segment. Here, every segment is linked to at least one mora and when two segments are linked to the same mora as in (6), this is taken to mean that they are produced simultaneously, so that the representation in (6) is interpreted as corresponding to a three-dimensional object, something like:

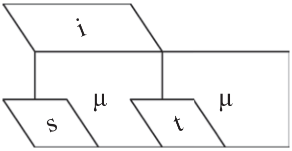
(7)



In this example, these C and V segments are very similar to traditional phones, but as will be apparent from the previous section, the C and the V segments are in fact intended to represent more complex segments, which are the basic building blocks of this framework.

The representation corresponding to (5b), then, would simply be:

(8)

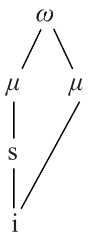


The similarity of the representations for words like “city” (6) and “sit” (8) is deliberate. In many languages, there are alternations between final consonants and final syllables with schwa. In French, for example, a word like “cite” is pronounced /sit/ with a final consonant in most (standard) dialects, but as /sitə/ in midi French with a very variable final schwa. Similarly, in Moroccan Arabic, Benkirane (1998) reports that:

the sequence CVC should be treated as disyllabic with the final consonant constituting the onset of a syllable with an empty rime. (...) a word such as /dib/ can be realised phonetically either with a final schwa [dibə] or with a final consonant which is devoiced but unquestionably released [dib<]. Both the pronunciation of a schwa and a release of the final consonant fulfil the same purpose of detaching the consonant from the preceding syllable. [p. 347]

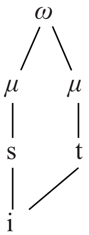
As in Moraic Phonology, multiple linking of a vowel or consonant can be used to indicate gemination or lengthening, so we can have a long vowel as in “see”:

(9)



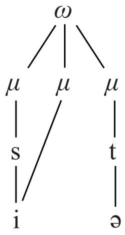
When the long vowel is followed by a consonant, the consonant is co-produced with the "preceding" vowel, that is, the end of the consonant is synchronised with the end of that vowel:

(10)



So in (10), the vowel is synchronised with the beginning of /s/ and with the end of /t/. This second consonant can, of course itself, be synchronised with a following vowel such as /ə/. So in "seater", for example, we have:

(11)



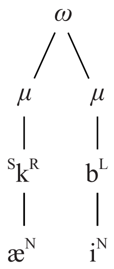
The whole question of the synchronisation of intervocalic consonants deserves attention — the question is obviously linked to the concept of ambisyllabic consonants, a concept which many linguists have found attractive, but which has had very little empirical backing.

In (6), there are two consecutive moras, so each consonant is attached to one of the two moras. But in (8), there are also two moras, and each consonant is attached to one of them.

In (1), however, the situation is more complicated, since there are still two moras but this time the first vowel /i/ is attached to both.

Returning to our example “scrambling”, we can now represent its prosodic structure of this word as:

(12)



With a sequence of two moras, each associated with a single complex C segment and V segment.

In these examples, I have not used the syllable as a prosodic constituent since the representation I have proposed removes one of the major motivations for the syllable as a prosodic unit. It remains an empirical question whether there may be other reasons to maintain the syllable as a prosodic constituent. In the rest of this paper, for convenience and for compatibility with previous work, I continue to use the syllable as a constituent intermediate between the mora and higher-level units.

4. Stress Feet

If we look above the level of the syllable, the correspondence between prosodic constituents and orthographic or syntactic constituents is no better.

Many descriptions of (British) English intonation and rhythm make use, following Abercrombie (1964) and Halliday (1967), of a unit called the *foot*, a concept originally proposed by Steele (1779) under the name of *cadence* or *bar*, and which is obviously derived from musical and/or poetical notation.

Since the term *foot* has also been used as a theoretical construct in metrical and autosegmental phonology, I shall use the term *stress foot* here for the unit proposed for the description of intonation. This unit can be defined for speech as a sequence of syllables[1] beginning with an accented syllable or with a silent beat at the beginning of a sentence, and continuing up to (but not including) the next accented syllable or silence.

With this definition, we can represent the sentence:

(13) They expected her election in September.

as:

(14) | ˆthey ex- | pected her e- | lection in Sep- | tember |

where the symbol [|] represents the foot boundary and [ˆ] a silent beat, similar to a pause in musical notation. As can be seen in this example, there can be a considerable mismatch between the level of syntactic words and that of stress feet.

Since Halliday (1967), most of the systematic descriptions of British English intonation [e.g. Crystal (1969), Cruttenden (1986), Tench (1996), up to and including Wells (2006)] have used, or implied, a framework, similar to this. The same phonological unit, the stress foot, is used in these studies to describe the (short-term) domains of both tone (melody) and quantity (rhythm).

Over 65 years ago, however, Jassem (1952) had suggested that we need *distinct* units to represent tonal and rhythmic structure. Jassem describes longer term stretches of intonation with a unit that he calls the *tune* or *tone group*. For the shorter units, he follows the practice of studies of tone languages and cites with approval Beach (1938), who says:

> In Chinese, for example, the syllable is universally accepted as the tone-unit, for the reason that practically every syllable of the language can mean different things according to the way it is intoned... In Panjabi and Lahuda, the tone-units are practically all disyllabic. In English and practically all other European languages, the tone-unit is neither the syllable, nor even the word, but a phrase consisting of one or more words. (p. 124).

Jassem adopts the term *Tonal Unit* rather than *Tone Unit*, presumably to avoid

[1] As mentioned at the end of Section 3, for convenience and compatibility with earlier work, I continue to refer to the syllable as the immediate constituent of higher level units. It would of course, be possible to represent the mora as directly linked to these larger units without using the syllable.

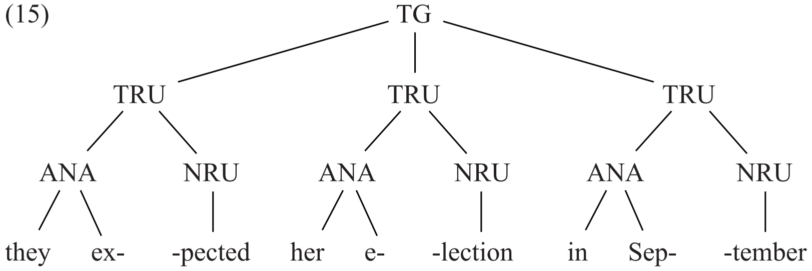
confusion between what he calls the *Tone Segment* in English and lexical tones in tone languages. It also helps to avoid confusion between this unit of prosodic structure and the longer term unit which he refers to as the *Tone Group*.

In fact, Jassem's definition of the *Tonal Unit*, given by a list of five different types of such units (op. cit. pp 49-50) is precisely equivalent to the definition given above of what Abercrombie, twelve years later, was to call the (stress) foot.

Unlike Abercrombie and Halliday, who use the same unit to describe both melody and rhythm, Jassem makes a clear distinction between the Tonal Unit, which is conceived of as the domain of occurrence of local pitch movements in English, and the *Narrow Rhythm Unit*, conceived of as the domain of segmental timing.

The *Narrow Rhythm Unit* is similar to the *Foot*, except for the fact that it does not usually cross word boundaries, except in cases of enclitics, which are treated prosodically as belonging to the previous word. Any syllables that are not parts of a *Narrow Rhythm Unit* (NRU), form an *Anacrusis* (ANA), in which, according to Jassem, the syllables are "pronounced extremely rapidly" (p. 40). The *Anacrusis*, together with the following *Narrow Rhythm Unit*, make up what Jassem termed the *Total Rhythm Unit* (TRU). Example (14) in this analysis would look like (15)。

(15)



The difference between the Narrow Rhythm Unit and the Anacrusis can be illustrated by a minimal pair, taken from Jassem (1949):

(16) *a*. summer dresses *b*. some addresses

In this example, he notes that although the phonemes and stresses are identical, there is a subtle difference of rhythm in the two, the first syllable of *summer* being shorter than that of *some*, whereas the second syllable of *summer* is shorter than the second syllable of *some a-*. He attributes this difference to the fact that the first two syllables of *summer* constitute a single Narrow Rhythm Unit, whereas in *some a-*, the first syllable constitutes a Narrow Rhythm Unit on its own and the second syllable constitutes an Anacrusis. He proposed to represent this in the phonetic transcription by the simple device of a space after each Narrow Rhythm Unit as in:

(17) a. / ˈsʌmə ˈdresiz/ b. / ˈsʌm əˈdresiz/

Here the spaces neatly correspond to the spaces in the orthographic transcription, but this is not always the case. Another example given by Scott (1940):

(18) a. Take Greater London b. Take Grey to London //

could be transcribed, using Jassem's proposal, as follows:

(19) a. /ˈteɪk ˈgreɪtə ˈlʌndn/ b. /ˈteɪk ˈgreɪ təˈlʌndn/

where the spaces are no longer identical to the orthographic version. In (20), the distinction between the two interpretations is not even reflected in the orthographic transcriptions:

(20) a. He bought her chocolates.
b. /hi ˈbɔ:thə ˈʧɒkləts/ (= for her)
c. /hi ˈbɔ:t həˈʧɒkləts/ (= belonging to her)

The definition of the Narrow Rhythm Unit, then, is a unit beginning with an accented syllable and ending before a rhythmic juncture. The rhythmic juncture corresponds in the majority of cases to the following word boundary except in the case of enclitics, which are assimilated to the preceding Narrow Rhythm Unit.

Klatt (1987), in his review of twenty years of research on speech synthesis, reached the pessimistic conclusion that:

> One of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena. (p. 760)

In a study of the segmental duration of a fairly large (5.5 hours) corpus of English, Hirst and Bouzon (2005) found that, as predicted by Jassem but contrary to Halliday's model, word boundaries *do* appear to play an important role in the rhythmic structure of English. Strong negative correlations were found between the duration of a segment and the number of phonemes in the stress foot, in the narrow rhythm unit and in the word, but no similar effect was found either in the syllable or in the anacrusis. Moreover, the correlation was greater for the NRU than either the stress foot or the word, thus confirming Jassem's predictions.

What is true for speech timing is also true for the study of tonal phenomena. As we saw above, Halliday used the same unit, the stress foot, to describe both pitch and rhythm; whereas in Jassem's model, these are dealt with by assuming different units for rhythm and for tone.

In a recent collection of articles, published to celebrate Wiktor Jassem's 90th birthday, I suggested (Hirst, 2012b) that Jassem's *Total Rhythm Unit* does not actually play any phonological role. Instead we can combine the *Tonal Units* and the *Rhythm Units* (*Anacruses* and *Narrow Rhythm Units*) into a single representation. In fact, if we do that, we can note that there is no longer any need to make a formal difference between *Anacrusis* and *Narrow Rhythm Unit*, since the *Narrow Rhythm Unit* will always coincide with the beginning of a *Tonal Unit* — both can simply be characterised as *Rhythm Units*.

We can note that a representation like this conforms to the *Strict Layer Hypothesis* (Selkirk, 1981, 2011), which states that:

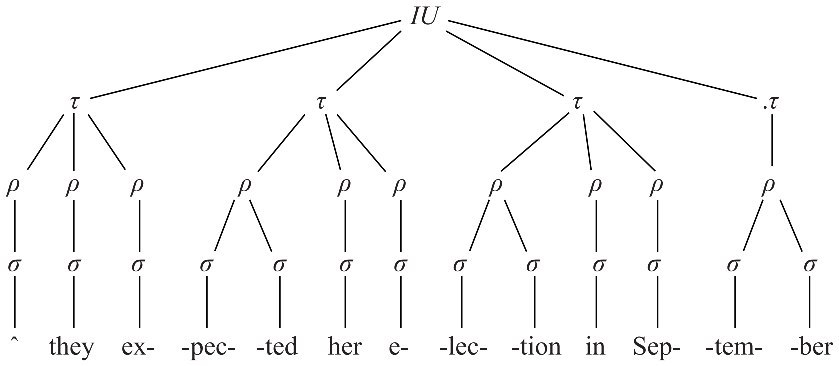
> A constituent of category-level n in the prosodic hierarchy immediately dominates only a (sequence of) constituents at category-level n-1 in the hierarchy. (Selkirk, 2011 p. 3.)

An empirical framework for the study of prosodic typology will obviously need a way to test which prosodic unit is the most effective in modelling the data. In ongoing work (Hirst, 2012a, 2015) on a prosody editor, designed specifically for linguists to test models of prosody, I consequently take a deliberately agnostic view on what constitutes the rhythm unit and what constitutes the tonal unit. Instead I *define* the *Rhythm Unit* [ρ] and the *Tonal Unit* [τ] as respectively the domains of interpretation of short-term planning of timing and pitch respectively and at the same time the *Intonation Unit* [IU] is *defined* as the domain of longer term interpretation of pitch and tone via changes in register and tempo.

This means that we can then formulate the differences between two hypotheses more precisely by saying, for example, that for Jassem, ρ corresponds to his *Narrow Rhythm Unit* and to his *Anacrusis*, whereas for Halliday, it corresponds to the *Stress Foot*. For Jassem, τ corresponds to Jassem's *Tonal Unit* which, as we saw, is identical to Abercrombie's and Halliday's *Stress Foot*.

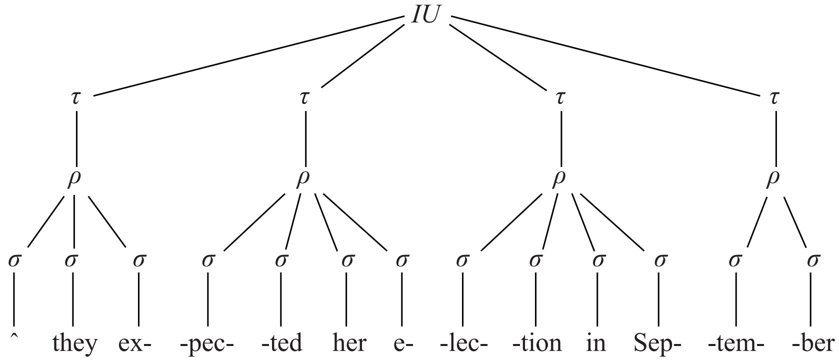
Using the annotation I propose, then, we can formulate a representation of example (14) in Jassem's model as:

(21)


IU
τ
τ
τ
.τ
ρ ρ ρ ρ ρ ρ ρ ρ ρ ρ
σ σ σ σ σ σ σ σ σ σ σ σ σ σ
ˆ they ex- -pec- -ted her e- -lec- -tion in Sep- -tem- -ber

Whereas in Halliday's model, the representation would be:

(22)


IU
τ
τ
τ
τ
ρ ρ ρ ρ
σ σ σ σ σ σ σ σ σ σ σ σ σ σ
ˆ they ex- -pec- -ted her e- -lec- -tion in Sep- -tem- -ber

where the *Rhythm Unit* and the *Tonal Unit* correspond to the same prosodic level.

It is an empirical question which of these two representations is the more appropriate for English. As mentioned above, our study of rhythm in English (Hirst & Bouzon op. cit.) concluded that Jassem's model (as in 21) provides a more adequate model than that of Halliday (as in 22).

5. Intonational Phrases

Higher level phonological categories present even worse problems of correspondence with the abstract syntactic structure. It is generally considered that syntactic phrases are recursive syntactic categories, whereas phonological phrases do not show the same type of recursive structure[1].

Since the relationship between phonological structure and syntactic structure seems far from direct, the solution adopted in most descriptions of intonation (particularly but not exclusively that of British English) has been to assume that the two types of structure are independent levels of representation and to suppose that there must exist some sort of *mapping rules* to explain how one type of structure is related to the other, even though the nature of these mapping rules is often far from explicit.

There is a fairly general consensus on the existence of a prosodic unit larger than the stress foot (= Tonal Unit) and shorter than the utterance. This is often called the *Intonational Phrase* in the autosegmental-metrical framework and corresponds to what was variously called the *Tone Group*, *Tune*, *Intonation Group* in earlier work on English intonation. As mentioned above, I use the term *Intonation Unit* for this prosodic constituent in line with the names of the lower level constituents, the *Tonal Unit* and the *Rhythm Unit*.

Crystal (1969) noted that the average length of Intonation Units in his corpus was of five words and that 80% of the Units were less than eight words long. When utterances are longer than this, they are usually broken up into two or more Intonation Units. Wh-questions appear to impose greater restrictions on the possible intonation breaks during the utterance so that a long question will still tend to be produced as a single Intonation Unit even in a sentence containing as many as eight accents:

(23) ˈWhat ˈmade ˈJohn ˈtell ˈAnne ˈnot to ˈgo ˈhome?

Apart from this type of question, it is fairly rare to find utterances in spontaneous speech which contain more than three or four accents in a single Intonation Unit.

[1] For arguments in favor of a (limited) form of recursivity in prosodic structure, see Ladd (1986).

There has been considerable disagreement as to what criteria, syntactic, semantic or pragmatic, are relevant for this phrasing. For a summary of arguments for and against syntactic constraints on phrasing cf. Couper-Kuhlen (1986), Chapter VIII.

Many of the arguments which have been presented against such constraints, however, no longer hold if we assume a less trivial correspondence between syntax and phonology than has generally been proposed.

Thus it has usually been assumed that a grammatical account of phrasing must show a one-to-one correspondence between syntactic units and prosodic units. This is obviously not the case in utterances like the following (from Couper-Kuhlen op. cit.) where [ ] indicate the boundaries of an intonation unit.

(24) a. [They feel like they're a forgotten bit] [of a war]
[that nobody wants to solve]
b. [They'll leave it alone] [till it splatters out] [to a deadly end]
c. [So here I am][in the middle of the most enormous][movement]
d. [as if the whole world] [is hanging waiting on our decision]
e. [which I found one of the most fascinating and most interesting] [times of my life]

The conclusion was consequently drawn that "it is virtually impossible to predict where boundaries will come." (p. 153)

I have suggested (Hirst, 1987, 1993) a different explanation for this apparent lack of correspondence between syntactic and phonological constituents. While pragmatic and phonological constraints are obviously the ultimate criteria by which a speaker decides *where* they will place a boundary, syntactic criteria define where these boundaries *may* occur.

In all the examples in (24), as well as in others given by the same author, it is striking that each boundary occurs before a complete syntactic constituent extending to the end of the sentence.

The reason why the correspondence between syntactic and prosodic constituents breaks down is that syntactic constituents may be interrupted by a prosodic boundary at the beginning of an internal syntactic constituent, but only if a prosodic boundary is also placed at the end of that constituent.

Thus in (24) for example, the syntactic structure relevant to the phrasings noted is:

(25) a. [They feel like they're a forgotten bit [of a war
[that nobody wants to solve]]]
b. [They'll leave it alone [till it splatters out [to a deadly end]]]
c. [So here I am [in the middle of the most enormous [movement]]]
d. [as if the whole world [is hanging waiting on our decision]]
e. [which I found one of the most fascinating and most interesting
[times of my life]]

This interpretation predicts that, while several different phrasings may be theoretically possible, others will be ruled out; in particular, internal boundaries are predicted not to occur before a constituent if the end of that constituent is not also marked by a boundary. In a sentence like:

(26) He promised to donate a considerable sum to her favorite charity.

Prosodic boundaries can occur after "donate" and after "sum", but the boundary after *donate* can only occur if there is also a boundary after *sum*.

6. Utterances

While there is quite a general consensus concerning the existence of prosodic constituents equivalent to what I have called *Intonation Units*, there is considerably less agreement as to whether larger or intermediate prosodic units need to be identified.

Several authors have proposed an intermediate constituent between the *Intonation Unit* and the *Stress Foot*. In ToBI annotation, for example (Silverman et al, 1992; Beckman et al, 2005), an *Intermediate Phrase* is distinguished from an *Intonational Phrase* by the fact that the former has only a final phrase accent, while the latter has both a phrase accent and a boundary tone. It is not obvious, however, that a sequence of two intermediate phrases necessarily form a higher-level constituent. An alternative would be to consider both as *Intonation Units* and to make the presence of a boundary tone an optional feature of this constituent.

It has been also suggested that *Intonation Units* are organized into higher-order *paratone-groups* (Fox, 1973, 1984) or *major paratones* (Yule, 1980), which are signalled essentially by a change of overall width of pitch range (Brazil, 1975; Brown, Currie & Kenworthy, 1980).

The beginning of a paratone is said to be usually marked by extra high pitch on the first accent, while the end is usually marked with extra-low pitch. When the end of a paratone is marked in this way but not the beginning, the result is what Yule has called a *minor paratone*.

It seems, however, equally possible to mark the beginning of a paratone but not the end. An alternative strategy would be to make a clear distinction between boundaries and constituents in the same way that I suggested above for the distinction between intermediate phrases and intonational phrases.

Rather than distinguish major and minor paratones, then, we might simply assume that Intonation Units can be marked as paratone-initial, paratone-final, or both or neither. Such a distinction could be marked in a transcription by simply doubling the initial or final square bracket of an intonation unit so that in a sequence:

(27) [[A] [B] [[C] [D]] [E]]

A and C are marked as paratone-initial and D and E as paratone final, even though the sequence as a whole is not properly bracketed (i.e. the number of opening brackets does not correspond to the number of closing brackets) and it cannot be divided into a sequence of independent paratones.

As Ladd noted in the quotation at the beginning of this presentation, theoretical ideas about constituent structure in phonology are definitely in need of considerable more exploration.

7. Conclusions

The framework I have sketched in this presentation makes a number of fairly controversial proposals.

The first is that the level of phones could be replaced by a more abstract level of contour segments, C and V such that the word "scrambling", for example, is composed of just 4 such segments.

The second is that work in the area of co-articulation suggests that, rather than stipulate that onset consonants are not connected to a mora but linked directly to the syllable, we could assume that the onset consonants are linked to the same mora as the "following" vowel and produced simultaneously. This proposal removes one of the major justifications for the syllable as a prosodic constituent.

The third suggestion, following the pioneering work of Wiktor Jassem, is that we should distinguish the domains of short-term planning of time and melody as Rhythm Units (ρ) and Tonal Units (τ), respectively.

The final suggestion is that making a clear distinction between the concepts of boundaries and that of prosodic constituents would mean that it is not necessary to assume higher level constituents other than the Intonation Unit.

In conclusion, the prosodic representation I propose is composed of 5 (or 6) levels: the complex segment, the mora, (the syllable), the rhythm unit, the tonal unit and the intonation unit. These constituents are organised into a strictly layered hierarchical structure as mentioned in section 4.

Of course, all these proposals are highly tentative and much further work and thought are needed on the subject of prosodic constituent structure.

References

Abercrombie, D. 1964. Syllable quantity and enclitics in English. In D. Abercrombie D. Fry, P. MacCarthy, N. Scott, J. Trim (eds.) *In Honour of Daniel Jones*, London: Longman, 216-222.

Allen, W.S. 1973. *Accent and Rhythm-Prosodic Features of Latin and Greek: A Study in Theory and Reconstruction*. Cambridge: Cambridge University Press.

Beach, D. 1938. *The Phonetics of the Hottentot Language*. Heffer and Sons., Cambridge, Mass.

Beckman, M.E. Hirschberg, J. & Shattuck-Hufnagel S. 2005. The original ToBI system and the evolution of the ToBI framework. In: S-A. Jun (ed.) *Prosodic Models and Transcription: Towards Prosodic Typology.*, London & New York: Oxford University Press, 9-54.

Benkirane, T. 1998. Intonation in Western Arabic (Morocco). In: D. J. Hirst & A. Di Cristo (eds.) *Intonation*

Systems: A Survey of Twenty Languages, Cambridge University Press, Chap 19, 348-362.

Brazil, David. 1975. Discourse analysis. *Discourse Analysis Monographs 1*. English Language Research, University of Birmingham.

Broselow, E. 1996. Skeletal positions and moras. In: J.A. Goldsmith (ed.) *The Handbook of Phonological Theory*., Blackwell Publishing.

Brown, G. Currie, K. & Kenworthy, J.. 1980. *Questions of Intonation*. London: Croom Helm.

Coleman, J. 2005. *Introducing Speech and Language Processing*. Cambridge: Cambridge University Press.

Couper-Kuhlen, E. 1986. *An Introduction to English Prosody*. London: Edward Arnold.

Cruttenden, A. 1986. *Intonation*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Fox, Anthony. 1973. Tone sequences in English. *Archivum Linguisticum*, 4:17-26.

Fox, Anthony. 1984. Subordinating and co-ordinating intonation structures in the articulation of discourse. In D. Gibbon & H. Richter (eds.), *Intonation, Accent and Rhythm. Studies in Discourse Phonology*, pp. 120-133. Walter de Gruyter, Berlin, 1984.

Halliday M. 1967. *Intonation and Grammar in British English*. Mouton.

Hayes, B. 1989. Compensatory lengthening in moraic phonology. *Linguistic Inquiry* 20(2): 253-306.

Hirst, D.J. 1985. Linearisation and the single segment hypothesis. In: J. Guéron, H.J. Obenauer, J.Y. Pollock (eds.) *Grammatical Representation*, Foris, Dordrecht: 87-100.

Hirst, D.J. 1987. *La représentation linguistique des systèmes prosodiques : une approche cognitive*. Thèse de Doctorat d'Etat (Habilitation Thesis), Université de Provence.

Hirst, D.J. 1993. Detaching intonational phrases from syntactic structure. *Linguistic Inquiry* 24(4): 781-788.

Hirst, D.J. 2012a. ProZed: A speech prosody analysis-by-synthesis tool for linguists. In: *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai.

Hirst, D.J. 2012b. Empirical models of tone, rhythm and intonation for the analysis of speech prosody. In: D. Gibbon, D. J. Hirst & N. Campbell (eds.) *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem*. Speech and Language Technology, Vol. 14/15, Polish Phonetic Association, Poznan: 23-33 .

Hirst, D.J. 2015. ProZed: A speech prosody editor for linguists, using analysis-by-synthesis. In: Hirose K, Tao J. (eds.) *Speech Prosody in Speech Synthesis: Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*., Springer Verlag, Berlin Heidelberg, Chap 1: 3-17.

Hirst, D.J., & Bouzon, C. .2005. The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). In: *Proceedings of Interspeech/Eurospeech* 05., Lisbon: 29-32.

Hyman, L.M. 1985. *A Theory of Phonological Weight*. Foris Publications.

Jassem, W. 1949. Indication of speech rhythm in the transcription of educated southern English. [in phonetic script]. *Le Maître Phonétique* III(92): 22-24.

Jassem, W. 1952. *Intonation of Conversational English*: (educated southern British). Nakl. Wroclawskiego Tow. Naukowego; skl. gl.: Dom Ksiazki.

Klatt, D. 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 82: 737-793.

Ladd, D.R. 1986. Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook* 3:

311-340.

Ladd, D.R. 2014. *Simultaneous Structure in Phonology*. Oxford: Oxford University Press.

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge: Cambridge University Press.

Ӧhman, S.E.G. 1966. Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America,* 39: 151-68.

Sagey, E.C. 1986. *The Representation of Features and Relations in Non-linear Phonology*. PhD thesis, Massachusetts Institute of Technology.

Scott, N. 1940. Distinctive rhythm. *Le Maître Phonétique,* 49: 6-7.

Selkirk, E. 1981. On prosodic structure and its relation to syntactic structure. In Thorstein Fretheim (ed.) *Nordic Prosody II: Papers from a Symposium*, Trondheim, TAPIR: 111-140.

Selkirk, E. 2011. The syntax-phonology interface. In J. Goldsmith, J. Riggle & A. C. L. Yu (eds.) *The Handbook of Phonological Theory 2*, Oxford: Blackwell: 435-484.

Silverman, K. Beckman, M. Pitrelli, J. Ostendorf, M. Wightman, C. Price, P. Pier- rehumbert, J. & Hirschberg. J. 1992. TOBI: A standard for labeling English prosody. In: *Second International Conference on Spoken Language Processing*, ISCA, Banff. Canada: 867-870.

Steele, J. 1779. *Prosodia Rationalis: An Essay towards Establishing the Melody and Measure of Speech, to be Expressed and Perpetuated by Peculiar Symbols*. (2nd ed.) J. Nichols, London.

Tench, P. 1996. *The Intonation Systems of English*. Cassell.

Wells, J. C. 2006. *English Intonation: An Introduction*. Cambridge: Cambridge University Press.

Yule, G. 1980. Speakers' topics and major paratones. *Lingua Amsterdam*, 52(1-2):33-47.

关于韵律结构的一些思考

赫　丹

法国国家科研中心语言与言语实验室

法国普罗旺斯地区艾克斯市艾克斯－马赛大学

摘　要　我们对韵律表征的想法很大程度上受到我们书面语言知识的影响。所有书面语言系统都将话语表征为由一个有限字符集中获取的线性成分序列。很多语言都将空格、标点符号之类的特殊字符用作边界符号。今天人们普遍认为，话语本身虽（分别）是以生理、声学和感知事件流产生、传播和感知的，但在大脑中却被表征为一种韵律结构，其中小的语音块按照音系层级结构组成更大的语音块，而且这种层次结构只与更为抽象的句法结构部分相关。本文列出并讨论了有关这些韵律块的性质以及韵律结构何以不同于书面语言和句法结构的一些想法，特别提出采用一种不太线性的方法研究韵律结构，可能会对韵律表征的性质产生重要有时甚至是令人惊讶的见解。

关键词　韵律结构　音子　韵素　音节　韵律　旋律　语调

Hirst, Daniel

Laboratoire Parole et Langage, CNRS & Aix-Marseille University

Aix-en-Provence, France

daniel.hirst@lpl-aix.fr