Syllable Parser Overview

H. Andrew Black
SIL International
7 November 2015

1 Introduction

The notion of a syllable has led to many linguistic analytical insights. This paper describes some key considerations that a syllable parser should meet. Besides being useful in exploring the nature of syllables in a given language, a syllable parser can be used to insert discretionary hyphens at appropriate syllable boundaries within a word and thus be useful in preparing typeset text.

2 General considerations

A syllable parser must take a word string as input, parse the string into its constituent segments and then produce the corresponding sequence of syllable structures, including showing which segments belong to which syllable. It also must be able to take a previously syllabified structure (which has since been modified by some kind of insertion or deletion of segmental material) and modify the syllable structure in an appropriate fashion to reflect the modification. That is, a syllable parser must be able to perform both syllabification and re-syllabification.

Currently this syllable parser only does initial syllabification.

2.1 Segment parsing

Given an input word string, the first task a syllable parser must perform is to break the input string into its constituent segments. Another way of looking at this is to say that there needs to be a mapping from the practical orthography to phonologically significant units.

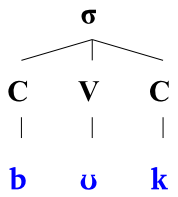
2.2 Syllabification overview

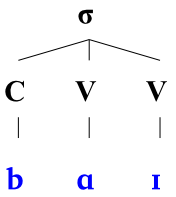
Based on the list of segments, a syllable parser must then build the syllables. At least four views of syllable internal structure have been proposed as discussed below. A syllable parser must be able to model at least one of these. Ideally, a syllable parser would have the flexibility to allow a user to model any one or more

of these four. Users then have the ability to model the view most familiar to them, to explore the advantages of the various views for their language, or to begin with a more simple model and incrementally move up the scale of complexity as they discover problems with the simpler model.

To illustrate the four views, consider how the English words *book* and *buy* would be syllabified. For this and all other English examples, we will represent the word using IPA script.[1] The four views are CV Patterns (Rensch n.d) as illustrated in (1), ONC Patterns (Pike & Pike 1947) as illustrated in (2), the Moraic approach (McCarthy & Prince 1986 and Hayes 1989) as in (3), and what I'll call the Nuclear Projection approach (Levin 1985) as in (4).

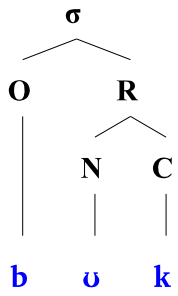
(1) CVC:

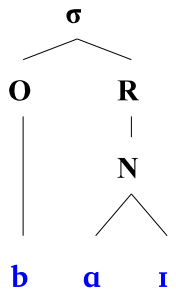
a. 

b. 

In (1), the syllable consists simply of a sequence of consonants and vowels. The user would supply the list of possible CV patterns and an indication of which segments were consonants and which were vowels. The program would seek to match the consonant vowel sequences of the input word against the possible CV patterns to exhaustively parse the word into syllables. It can do this by a left-to-right sweep of the word.[2]

(2) ONC:

a. 

b. 

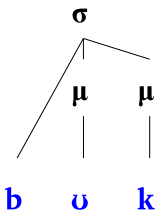
In (2), there is more structure to the syllable. There is an obligatory nucleus (N). This nucleus forms a rime (R) together with an optional coda constituent (C). The coda may consist of one or more consonants. An optional initial onset element (O) also can exist. Like the coda, it may consist of one or more consonants. The

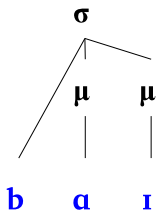
[1]Throughout this document, the following symbols are used: σ = syllable, IO = onset, R = rime, N = nucleus, C = coda, and μ = mora.

[2]We realize that there are Non-Roman scripts which are not left-to-right in orientation. Even in these situations, we suspect that it is the case that these can be encoded in a left-to-right roman script. All processing would then be done in this roman script.

nucleus may consist of more than one vowel as shown in (2b) for the word *buy* (baɪ). The program must seek to match onset, nucleus and coda symbols as part of its attempt to exhaustively parse the word into syllables. It can do this by a left-to-right sweep of the word.

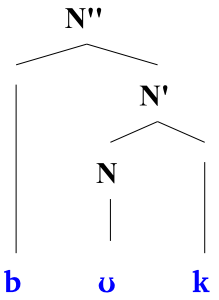
(3) Moraic:

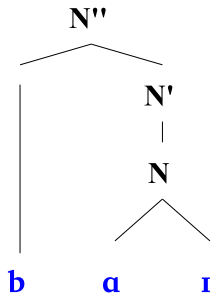
a. 

b. 

In (3), issues of syllable weight are considered. Many languages have word stress patterns that make a distinction between light and heavy syllables. The Moraic approach seeks to model that distinction via a unit of syllabic weight called the mora (symbolized in (3) by the Greek symbol mu - μ). Light syllables have one mora while heavy syllables have two.[3] The program must parse the segments of the word in such a fashion that it can determine which ones should be moraic and which ones should be onsets. It can do this by a left-to-right sweep of the word.

(4) Nuclear-projection:

a. 

b. 

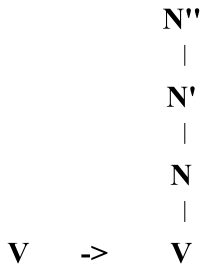
The next approach to syllabification we will consider involves building constituents in a predetermined fashion. Patterned after some approaches to syntactic constituent structure, Levin (1985) proposes what we will call the Nuclear-projection approach. Noting that every syllable has a nucleus, this approach projects it as the "category" `N`. It has a "complement" (or right sister) of a coda, dominated by the `N'` (pronounced N-bar) first level projection. The onset is the "specifier" of the head `N` and is thus dominated by the second level projection `N''` (pronounced N-double-bar).

Unlike the other three approaches, a single left-to-right sweep of the word is not appropriate for the Nuclear-projection approach. These X-bar-like structures are built by at least the three rules shown in (5). Each rule is applied in a separate

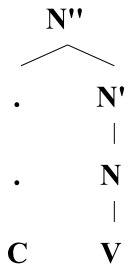
[3] One may use the Moraic approach for a language that only has light syllables by setting a "maximum number of moras per syllable" parameter to 1.

sweep of the word. First, all vowels project an N and associated N' and N'' constituents (5a). The rule in (5b) then assigns a pre-vocalic consonant to the onset position. Finally, rule (5c) assigns any post-vocalic consonant not already in syllable structure to coda position. These three rules are applied in the order given.

(5) a.



b.



c.

Another approach to syllabification especially popular among generative linguists today is the one used in Optimality Theory (Prince & Smolensky 1993). We will call it the OT approach. For our purposes here, we will essentially follow the tack taken by Hammond (1997) for syllable parsing. While Hammond uses just onsets, nuclei, and codas, we will augment it slightly to also include rimes. Thus, the syllable internal structure will be identical to the ONC approach. The algorithm, though, for parsing syllables will be quite different from the ONC approach.

2.3 Other issues in syllabification

There are some other issues that need to be kept in mind as well as one considers syllabification.

2.3.1 Codas

First, not every language allows codas (although most do). Thus, there needs to be a parameter indicating whether or not codas even exist in the language being syllabified.[4]

2.3.2 Sonority Sequencing Principle (SSP)

All approaches except for the CV Pattern one must address an additional issue (at least as far as these requirements are concerned). Consider a form like *frantic* (fɹæntɪk). Given that more than one consonant may appear in both the onset and the coda constituents, how should the nt consonant sequence be treated? Are both consonants in the coda of the first syllable or are both in the onset of the second syllable or should the first be in the coda of one syllable and the other be in the onset of the next syllable and how do we know?

For over a century, linguists have realized that onset and coda construction is governed by the Sonority Sequencing Principle (SSP) (see Kenstowicz 1994:254-255 for a discussion). The SSP states that onsets increase in sonority as they approach the nucleus and codas decrease in sonority as they get further from the

[4]The OT approach handles this via a NoCoda constraint.

nucleus. The scale of sonorancy is generally agreed to be as shown in (6), where vowels are more sonorous than glides, etc.

(6)

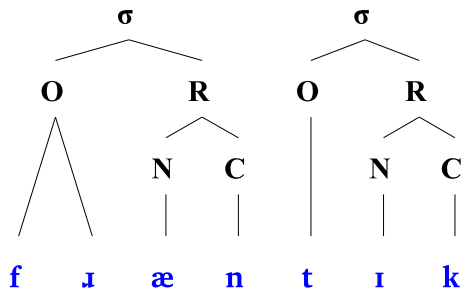
vowels	*more sonorous*
glides	
liquids	
nasals	
obstuents	*less sonorous*

Given the SSP, we know that the nt sequence cannot form an onset sequence since n (a nasal) is more sonorous than t (an obstruent). The SSP, however, does not definitively determine the syllable constituency. Since n is more sonorous than t, both could be in the coda of the middle syllable. It is also possible for there to be a syllable break between the n and t, thus putting the n in the coda of the first syllable and the t in the onset of the final syllable. Which is to be preferred?

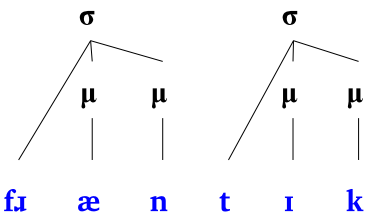
2.3.3 Onset Principle

There is a general tendency in languages to avoid having onsetless syllables which Itô (1989:223) refers to as the Onset Principle.[5] Since English follows this tendency, the n should be in the coda of one syllable and the t should be in the onset of the last syllable as shown in (7) for each of the three pertinent approaches.

(7) a.

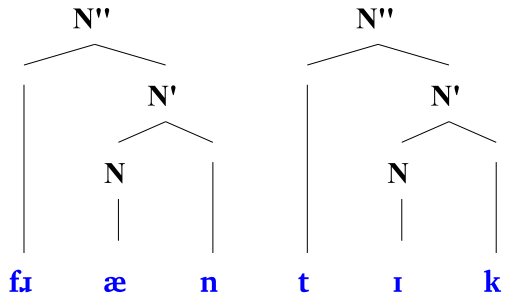


b.



[5]Many languages which require an onset word medially do not also require the initial syllable to have onsets.

c.



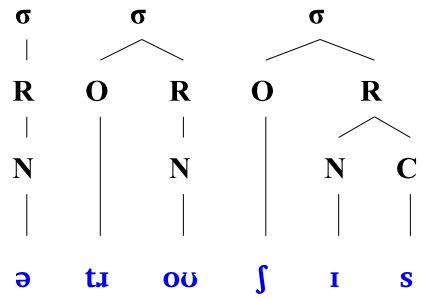
Note, however, that not all languages use the Onset Principle. Therefore, we must treat it as a parameter (even though we will continue to call it a principle). Since some languages require all but the first syllable to have an onset, this parameter has three values:[6]

- every syllable must have an onset,
- all but the first syllable must have an onset,
- onsets are not required.

2.3.4 Onset Maximization

Now consider the word *atrocious* (ətɹoʊʃɪs). How should the tɹ sequence be handled? It cannot be in the coda of the first syllable because t (an obstruent) is less sonorous than ɹ (a liquid). Since there is rising sonority between the two consonants, the tɹ sequence could constitute the onset of the second syllable as shown[7] in (8).

(8)

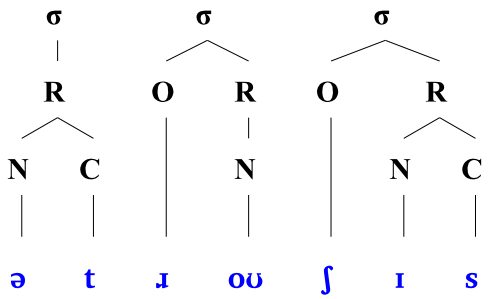


Alternatively, the sequence could be split between the two syllables: t would be in the coda of one syllable and ɹ would be in the onset of the other as shown in (9).

[6]The OT approach uses two constraints to handle this parameter: ONSETWORDINITIAL and ONSETWORDMEDIAL.

[7]I use the ONC approach for illustrative purposes, but the same would apply for the Moraic and the Nuclear-projection approaches

(9)

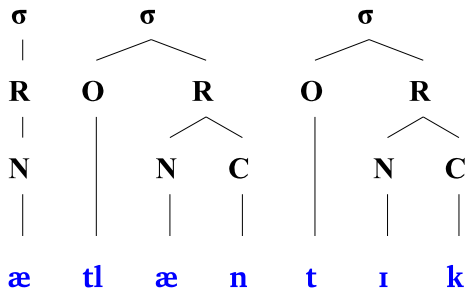


For English, example (8) is the correct structure. That is, like many other languages, English maximizes its onsets: if the consonants in question are increasing in sonority, they are all placed in the onset. Since not every language maximizes its onsets, we need a parameter of Onset Maximization.[8] So far, then, we have one principle (the SSP) and three parameters (Codas, the Onset Principle and Onset Maximization).

2.3.5 Filters

Now consider the word *Atlantic* (ætlæntɪk). Like frantic, *Atlantic* also has a word medial nt sequence. Notice, however, that it also has a tl consonant sequence. Given the SSP and Onset Maximization, we would expect this word to be syllabified as in (10) where the tl sequence constitutes the onset of the second syllable. Recall that t (an obstruent) is less sonorous than l (a liquid). Thus, tl is a perfectly fine onset as far as the SSP is concerned.

(10)

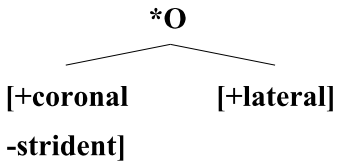


English, however, never allows such a tl sequence in onset position. One solution to this problem is to posit *filters* to rule out such sequences (Clements & Keyser 1983).[9] Such a filter is shown in (11) where all `+coronal, -strident` segments are not allowed in an onset when they are immediately followed by a lateral.

[8]The OT approach deals with Onset Maximization by an appropriate ranking of the onset *COMPLEX and NOCODA constraints.

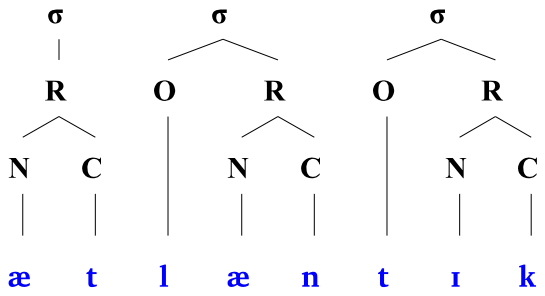
[9]The OT approach does not employ filters.

(11)



Given such a filter, the syllabification will be as in (12).

(12)



2.3.6 Templates

2.3.6.1 Motivation for templates

Now consider a word like *strap* (stɹæp). The SSP allows the tɹ sequence to be in the onset position, but what about the st sequence? It, too, is in the onset, yet here we have a violation of the SSP since s is at least equal in sonority if not actually more sonorous than t. This happens not only word initially, but also word medially as a form such as *monster* (mɑnstəɹ) demonstrates.

English has some other SSP violations. Consider the coda of words such as *fifth* (fɪfθ), *sixth* (sɪksθ) or even *sixths* (sɪksθs). These unusual forms have sequences of one, two or even three final coronal consonants all of which violate the SSP.

While a number of approaches have been posited to deal with this abnormal behavior (see Kenstowicz 1994:258-261 for a discussion), as far as this paper is concerned we can treat them all by adding a mechanism of *templates*.[10] We'll allow for two major kinds of templates and call them constituent templates and word templates. Constituent templates can be used for the English *s* case while word templates can be used for the English cases like *sixths*. Constituent templates deal with issues relating to onset, rime, nucleus, coda or syllable constituents. That is, they are for the syllable constituent or any of its sub-constituents. Word templates, on the other hand, deal with special conditions that occur only word initially or word finally.

[10]The OT approach does not employ templates.

2.3.6.2 The nature of templates

Each such template (whether word or constituent) consists of one or more positions, each of which may be tagged as optional and each of which may be subject to the SSP or not. Each position may further be limited to a particular set of segments by indicating their appropriate natural class. Thus we can handle the English s problem by using an onset-oriented constituent template like the one in (13).[11]

(13) Onset Template:

s	[voicelessNonCont]	[sonorantCV]
-ssp		

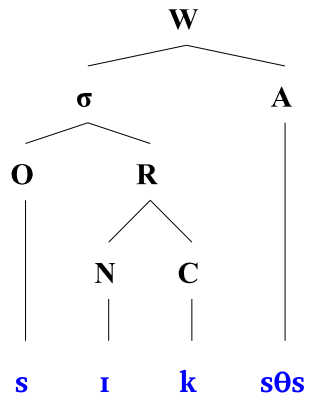
The final coronal sequences can be handled via a word template like the one in (14), where `-ssp` means an SSP exception and `opt` means the position is optional.

(14) Word Final Template:

[coronal]	[coronal]	[coronal]
-ssp	-ssp	-ssp
	opt	opt

A word final template like the one in (14) is treated as an appendix to the end of the word (similarly, a word-initial template is treated as an appendix to the beginning of the word). Thus, sɪksθs has the structure shown in (15), where the `W` constituent is the word and the `A` constituent is an appendix.

(15)



[11] As Kenstowicz (1994:258) notes, if one is using the Nuclear-projection approach, one can also write a rule to deal with the special status of the s as shown in (i).

(i)

Some other observations about templates are in order.

The only time the parser should attempt to add an appendix to a word is when the parser was unable to successfully syllabify the entire word. This reflects the exceptional nature of word templates. Further, any appendix should be minimal: the parser should take what syllable structure it was able to build and then add any appendix structure needed to account for the remaining segments.

Constituent templates, on the other hand, are always tried while building the appropriate constituent. In fact, the tack taken here is that if there is at least one constituent template for a particular constituent, then at least one of those constituent templates must be met in order for that constituent to be valid. We note that constituent templates only make sense for the ONC and Moraic approaches. The Nuclear-projection approach will use a rule instead of a constituent template.

2.3.6.3 A templatic approach to syllabification

The fact that we will treat constituent templates as obligatory has special implications for the use of a constituent template which is for the entire syllable. If the constituent template is for the entire syllable, then every syllable must match the template and meeting the template becomes the dominating factor in the syllabification. This allows for approaches to syllabification like the one proposed in Itô (1989) which has a required template for the syllable constituent.

For example, we could posit a constituent template such as the one in (16).

(16) Syllable template:

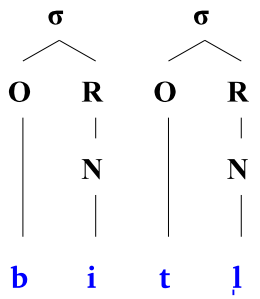
[C]	[V]	[V]	[Nasal]
opt		opt	opt

This effectively says that syllables in this language consist of any of the following patterns (where N indicates a nasal): V, VV, VN, VVN, CV, CVV, CVN, CVVN. In some ways, then, it is like the CV approach. There are at least two crucial distinctions, however. One is that the syllable constituency of the ONC or Moraic approach is still maintained. The other is that unlike the CV approach, one can still have the Sonority Sequencing Principle apply for complex onsets and/or codas.

2.3.7 Syllabic consonants

Now consider a form like *beatle* (bitl̩), where there are two syllables, but only one vowel. In the second syllable, the l̩ is the nucleus of the syllable – it is a "syllabic consonant". Thus, this word would be syllabified as in (17).

(17)

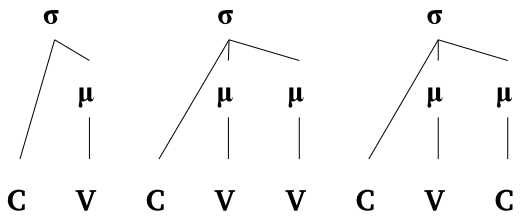


The syllable parser will need to be able to handle such syllabic consonants.

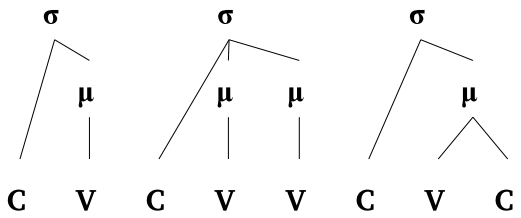
2.3.8 Weight by Position

Finally, the Moraic approach has another consideration. Recall that this approach models the difference between light and heavy syllables by using one mora for light syllables and two moras for heavy syllables. It turns out that not every language which makes a distinction between heavy and light syllables treats syllables with codas (i.e. CVC) as being heavy. For instance, Hayes (1989:255) notes that while Latin treats both CVV and CVC syllables as heavy and CV syllables as light, Lardil treats only CVV syllables as heavy and treats both CVC and CV syllables as light.

(18)



(19)



To account for those languages which do treat codas as heavy, Hayes (1989:258) gives a rule of Weight by Position. As he says, "… the basic idea is that certain coda consonants are given a mora when they are adjoined to the syllable…".

2.3.9 OT Constraints

The OT approach uses a set of ranked and violable constraints. A set of potential candidate syllable parses is generated and then these candidates are evaluated

against the set of ranked and violable constraints. The candidates which emerge are posited as the correct parses (although given the nature of the constraints, it is extremely unusual for more than one parse to emerge). The set of constraints is considered to be universal. The claim is that the different rankings of these constraints produces the different kinds of syllabification patterns exhibited by the world's languages.

References

Clements, George N. and Samuel J. Keyser. 1983. *CV phonology: A generative theory of the syllable.* Cambridge, Mass: MIT Press.

Hammond, Michael. 1997. Parsing in OT. University of Arizona. Manuscript. ftp://ruccs.rutgers.edu/pub/OT/TEXTS/archive/222-1097/222-10972.rtf

Hayes, Bruce. 1989. Compensatory lengthening in moraic phonology. *Linguistic Inquiry* 20:253-306.

Itô, Junko. 1989. prosodic theory of epenthesis. *Natural Language and Linguistic Theory* 7:217-259.

Kenstowicz, Michael. 1994. *Phonology in Generative Grammar.* Cambridge, Mass: Basil Blackwell.

Levin, Juliette. 1985. *metrical theory of syllabicity.* Ph.D. dissertation. Cambridge, Mass. MIT.

McCarthy, John and Alan Prince. 1986. Prosodic morphology. Waltham, Mass: Brandeis University. Manuscript.

Pike, Kenneth L. and Eunice Pike. 1947. Immediate constituents of Mazateco syllables. *International Journal of American Linguistics* 13:78-91.

Prince, Alan and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Rutgers University. Manuscript.

Rensch, Carolyn M. n.d. *Problems for introduction to phonology, Part II – Handouts.* Dallas, TX: Summer Institute of Linguistics.