`syllabify.py`: Automated English syllabification

Kyle Gorman

October 27, 2012

This document describes the principles of English syllabification implemented by `syllabify.py`, a Python module for assigning prosodic structure to ARPABET transcriptions.[1]

1 Ambiguous segments

The syllabification procedure begins by separating sequences of vocalic and consonantal segments. In English, *r* and onglides pattern with consonants or with vowels depending on the context in which they occur. The heuristic adopted here is that ambiguous segments which impose restrictions on adjacent vowels are themselves vocalic, and those which impose restrictions on adjacent consonants are consonantal.

Initially, between two vowels, or finally, *r* is consonantal. Before another consonant, however, *r* has been lost in Received Pronunciation. Even in *r*-ful dialects, though, post-vocalic non-onset *r* patterns with vowels, not coda consonants. Before non-onset *r* many vowel contrasts are suspended (e.g., Fudge 1969:269f., Harris 1994:255): compare American English *fern*/*fir*/*fur* to *pet*/*pit*/*putt*. In this position, *r* is the only consonant which permits variable glottalization of a following /t/ in *r*-ful British dialects (Harris 1994:258), and the only consonant which does does not trigger variable deletion of a following word-final /t, d/ in American dialects (Guy 1980:8). This is shown in (1–2) below.

(1) /t/-Glottalization in *r*-ful British dialects:

a.	des[ɚt]	∼	des[ɚʔ]
	c[ɚt]ain	∼	c[ɚʔ]ain
b.	fi[st]	∼	*fi[sʔ]
	mi[st]er	∼	*mi[sʔ]er

(2) /t, d/-Deletion in American English:

a.	be[lt]	∼	be[l]
	me[nd]	∼	me[n]
b.	sk[ɚt]	∼	*sk[ɚ]
	th[ɚd]	∼	*th[ɚ]

Consequently, pre-consonantal *r* is assigned to the preceding nucleus.

[1] The code is available at GitHub (URL: `http://github.com/kylebgorman/syllabify`).

The front onglide is assigned to onset position when initial or preceded by a single consonant, as in [j]*arn* or *ju*[n.j]*or*. When the glide is preceded by two or more consonants, it is assigned to the nucleus. There is considerable evidence in support of this assumption. When [j] is assigned to the onset, it may be followed by any vowel (Borowsky 1986:276), but when it is nuclear, the following vowel is always [u] (Harris 1994:61f., Hayes 1980:232). Clements and Keyser (1983:42) note that [j] is the only consonant which can follow onset /m/ and /v/: [mj]*use*, [vj]*iew*. Finally, [ju] sequences in words such as *spew* pattern together in language games (Davis and Hammond 1995, Nevins and Vaux 2003) and speech errors (Shattuck-Hufnagel 1986:130).[2]

The phonotactic properties of the back onglide [w] are quite different than those of the front onglide, and it is consequently assigned to the onset portion of medial clusters. Whereas [j] shows only limited selectivity for preceding tautosyllabic consonants (Kaye 1996), [w] only rarely occurs after onset consonants other than [k] (e.g., *tran*[kw]*il*), and never after tautosyllabic labials in the native vocabulary. Whereas [kj] is always followed by [u], [kw] may precede nearly any vowel (Davis and Hammond 1995:161).

2 Parsing medial consonant clusters

Medial consonant clusters are segmented into coda and onset using a heuristic version of the principle of onset maximization (e.g., Kahn 1976:42f., Kuryłowicz 1948, Pulgram 1970:75, Selkirk 1982:358f.) which favors parses of word-medial clusters in which as much of the cluster as possible is assigned to the onset. A medial onset is defined to be "possible" simply if it occurs word-initially (according to the rules defined above). As an example, the medial clusters in words such as *neu*[.tɹ]*on* or *bi*[.stɹ]*o* also occur in word-initial position (e.g., [tɹ]*ain*, [stɹ]*ike*), so the entire cluster is assigned ot the onset. In contrast, the cluster in *mi*[n.stɹ]*el* is not found word-initially; the maximal onset here is [stɹ] and the remaining [n] is assigned to the preceding coda.

In English, when a medial consonant cluster is preceded by a stressed lax vowel, as *wh*[ɪs.p]*er*, *v*[ɛs.t]*ige*, or *m*[ʌs.k]*et*, the first consonant of the cluster checks the lax vowel (Hammond 1997:3). As Harris (1994:55) notes, however, when the medial cluster is also a valid onset, as in *whi*[s.p]*er*, *ve*[s.ti]*ge*, and *mu*[s.k]*et*, onset maximization will incorrectly assign the entire cluster to the onset and leave the lax vowel unchecked. For this reason, onset maximization parses are modified to assign the first consonant of a complex medial consonant cluster to the coda before a stressed lax vowel (Pulgram 1970:48).

3 Phonologization

The traditional analysis of affricates as single segments (e.g., Chomsky and Halle 1968:321f., Jakobson et al. 1961:24) rather than stop-fricative sequences (e.g., Hualde 1988, Lombardi 1990) is assumed. In many languages, affricates pattern with simple onsets; for instance, Classical Nahua bans true onset clusters but permits the affricate series [ts, tʃ, tɬ] (Launey 2011:9). In English, however, affricates do not form complex onsets. Yet, the stop and release phase of affricates cannot be

[2]The glide is also assumed to be present in underlying representation (e.g., Anderson 1988, Borowsky 1986:278) rather than inserted by rule (e.g., Chomsky and Halle 1968:196, Halle and Mohanan 1985:89, McMahon 1990:217) since presence or absense of the glide is contrastive (e.g., *booty*/*beauty*, *coot*/*cute*).

separated by a syllable boundary, as predicted from the assumption they are single phonological units.

In English, [ŋ] has been analyzed as a pure allophone of /n/ before underlying /k, g/ (with later deletion of /g/ in some contexts; Borowsky 1986:65f., *SPE*:85, Halle and Mohanan 1985:62), or as a phoneme in its own right (e.g., Jusczyk et al. 2002, Sapir 1925). Onset [ŋ] is totally absent in onset position, where it cannot be followed by a /k, g/ needed to derive the velar allophone, a fact predicted only by the former account, and English speakers have considerable difficulty producing initial [ŋ] (Rusaw and Cole 2009). The allophonic analysis is assumed and surface [ŋ] is mapped to underlying /n/.

4 Using `syllabify.py`

This section assumes some knowledge of the Python programming environment, and installation of a current version (a version from the 2.7 branch). `syllabify.py` is a Python *module*, and to use it, you must write Python code to interface with it. The following is a simple example from the author's research (the code is also included in the file `kp.py` distributed with this package).

Pierrehumbert (1994:175f.) claims that in English, medial codas ending in /k, g/ are rarely followed by a labial onset (i.e., /p, b, f, v/). To begin to evaluate this claim, we search for clusters that consist of this structure in the CMU pronunciation dictionary.[3] Each entry is syllabified, and the orthographic form of entries that match the structural description are printed to standard output.

```
from syllabify import syllabify

if __name__ == '__main__':
    source = open('cmudict.0.7a', 'r')
    for line in source:
        if line[0] == ';':                  # header, comments
            continue                        # starts next iter of loop
        (word, pron) = line.rstrip().split('  ', 1)
        syllables = syllabify(pron.split())  # syllabify pronunciation
        for i in xrange(len(syllables) - 1):
            coda  = syllables[i][2]
            onset = syllables[i + 1][0]
            if coda and coda[-1] in {'K', 'G',}:
                if onset and onset[0] in {'P', 'B', 'F', 'V',}:
                    print word
                    break                   # don't print a word twice
```

If you place this in your working directory in a file called `kp.py`, you can call it simply by running `python kp.py`. Assuming you also have copies of `cmudict.0.7a` and `syllabify.py` in you working directory, this will produce output like the following:

[3] You will need to download a recent version (URL: `https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict/cmudict.0.7a`) and place it in your working directory.

```
AGFA
AKBAR
AKBAR(1)
AKBASH
BACKBITE
BACKBITING
BACKBOARD
BACKBOARDS
BACKBONE
BACKBONES
...
```

As can be seen, exceptions occur, but many are complex words.

References

Anderson, John M. 1988. More on slips and syllable structure. *Phonology* 5:157–159.

Borowsky, Toni. 1986. Topics in the lexical phonology of English. Doctoral dissertation, University of Massachusetts, Amherst. Published by Garland, New York, 1991.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English.* Cambridge: MIT Press.

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

Davis, Stuart, and Michael Hammond. 1995. On the status of onglides in American English. *Phonology* 12:159–182.

Fudge, Erik C. 1969. Syllables. *Journal of Linguistics* 5:253–286.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Locating language in time and space*, ed. William Labov, 1–35. New York: Academic Press.

Halle, Morris, and K. P. Mohanan. 1985. Segmental phonology of Modern English. *Linguistic Inquiry* 16:57–116.

Hammond, Michael. 1997. Vowel quantity and syllabification in English. *Language* 73:1–17.

Harris, John. 1994. *English sound structure.* Cambridge: Blackwell.

Hayes, Bruce. 1980. A metrical theory of stress rules. Doctoral dissertation, MIT.

Hualde, Jose Ignacio. 1988. Affricates are not contour segments. In *Proceedings of the 7th West Coast Conference on Formal Linguistics*, 143–157. Stanford, CA: Stanford Linguistics Association.

Jakobson, Roman, Gunnar Fant, and Morris Halle. 1961. *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge: MIT Press.

Jusczyk, Peter W., Paul Smolensky, and Theresa Allocco. 2002. How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition* 10:31–37.

Kahn, Daniel. 1976. Syllable-based generalizations in English phonology. Doctoral dissertation, MIT. Published by Garland, New York, 1980.

Kaye, Jonathan. 1996. Do you believe in magic? the story of s+C sequences. In *A festschrift for Edmund Gussmann*, ed. Henryk Kardela and Bogdan Szymanek, 155–176. Lublin: Lublin University Press.

Kuryłowicz, Jerzy. 1948. Contribution à la théorie de la syllabe. *Bulletin de la Société Polonaise de Linguistique* 8:80–114.

Launey, Michel. 2011. *An introduction to Classical Nahuatl*. New York: Cambridge University Press.

Lombardi, Linda. 1990. The nonlinear organization of the affricate. *Natural Language and Linguistic Theory* 8:375–425.

McMahon, April. 1990. Vowel shifts, free rides and strict cyclicity. *Lingua* 80:197–225.

Nevins, Andrew, and Bert Vaux. 2003. Metalinguistic, shmetalinguistic: The phonology of shm-reduplication. In *Papers from the 39th meeting of the Chicago Linguistic Society*, 702–721. Chicago: Chicago Linguistic Society.

Pierrehumbert, Janet. 1994. Syllable structure and word structure: A study of triconsonantal clusters in English. In *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, ed. Patricia A. Keating, 168–188. Cambridge: Cambridge University Press.

Pulgram, Ernst. 1970. *Syllable, word, nexus, cursus*. The Hague: Mouton.

Rusaw, Erin, and Jennifer Cole. 2009. Learning constraints that oppose native phonotactics from brief experience. Paper presented at the Mid-Continental Workshop on Phonology.

Sapir, Edward. 1925. Sound patterns in language. *Language* 1:37–51.

Selkirk, Elisabeth O. 1982. The syllable. In *The structure of phonological representations*, ed. Harry van der Hulst and Norval Smith, 337–385. Dordrecht: Foris.

Shattuck-Hufnagel, Stefanie. 1986. The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. *Phonology Yearbook* 3:117–149.