Awesome short-a funtimes

Hilary Prichard (University of Pennsylvania)

Kyle Gorman (OHSU)

**Introduction**

**Short-a in Philadelphia**

In Early Modern English, short-a underwent irregular lengthening and tensing before the voiceless fricatives /f, θ, s/ and nasals /m, n/; reflexes of this pattern are present in Mid-Atlantic dialects of American English as well as in various British dialects (see Wells 1982: I.203f.). Labov (1981) argues that the distribution of the tense and lax phonemes in the Mid-Atlantic is no longer governed by allophonic rules, but rather is the product of a complex phonemic split. In addition to the existence of several minimal pairs (e.g., the lax modal auxiliary *can* vs. the tense noun *can*), Payne (1980) finds that children born out-of-state who move to the Philadelphia metropolitan area fail to acquire the local short-a distribution, though they participate fully in other regional sound changes, which are largely allophonic or phonologically unconditioned.

Ferguson (1972) provides one of the first descriptions of the rules underlying the traditional Philadelphia split short-a system. Ferguson's research was carried out between 1940 and 1960, and is based on his own native speaker intuition, supplemented by natural observation and judgments from informants, mostly young men who attended the University of Pennsylvania. Ferguson first defines the phonetic quality of the two phonemes which make up the split system, noting that (æ) is low, front, unrounded, monophthongal---essentially of the same quality as most other varieties of English---while (æh) by contrast is diphthongal, and can be raised to [ɛə] or even [eə]. As to the general rule governing which phoneme surfaces, Ferguson gives:

(1) The “traditional”pattern (after Ferguson 1972:262):  
  


Ferguson identifies numerous exceptions to this rule. For instance, there are two main categories of lexical items which have the lax phoneme rather than the tense phoneme predicted by (1): the verbs *am, can, ran, began, hast,* and *hath,* and the “learned words” such as *aft, crass, daft, damsel, gaffe, lass,* and *wrath* (264). Conversely, the “adjectives of emotion” *mad, bad,* and *glad* have the tense phoneme even though following [d] is not a regular tensing environment (263). In addition to these lexical exceptions, there are three classes of morphophonological exceptions. First, words like *passing* and *classes*, which fail to meet the conditions of (1) due to the addition of inflectional suffixes (the progressive/future/gerund suffix /-ɪŋ/, the noun plural suffix /-z/, the third-person singular active indicative agreement suffix /-z/, the regular past/past participle suffix /-d/, and perhaps the agentive   
/-ɚ/) are tense; Ferguson calls this the “preservation of stem identity”. On the other hand, tensing does not occur when the context in (1) is met due to shortening (*math*, *exam*) or schwa-elision (*family, camera*, *Catholic*)*.* Finally, according to Ferguson, short-a is not tense in an initial stressed syllable when followed by *-sp, -sf,* or *-sb* (*aspirin, asphalt, asbestos*).

Labov (1989, 1994) builds on Ferguson's description, formalizing additional rules and investigating possible exceptions using data from LCV, a corpus of sociolinguistic interviews collected in Philadelphia between 1973 and 1977. Whereas Ferguson's rule (1) requires that a tensing consonant be word-final or followed by another consonant, Labov specifically refers to a requirement that the tensing consonant is tautosyllabic (1994: 430). The result is that tensing is limited to closed syllables. He adds that the addition of inflectional suffixes does not affect the distribution of the tense and lax phonemes, whereas there is extensive variation associated with derivational suffixes.Labov also finds variation when short-a is followed by /st/ clusters (*master, plaster*), and observes possible extension of tensing to contexts where short-a is followed by nasal-vowel or lateral-vowel sequences such as *planet* or *personality* (1994: 433). He finds that the exceptional status of *mad, bad, glad* is truly lexical, citing the contrast of tense *bad* with lax *badminton*, and tense *glad* with lax *Gladstone* (1989: 17).

**Data/Methods**

Iterative approach to recoding. Started with the original variable class, identified big errors, came up with new coding scheme. Re-ran data, analyzed new variable class. Rinse, repeat. Similar strategy to Labov 1989 – number of tense/lax instances of individual words in variable/unclear environments are tabulated to determine category membership.

How do we measure whether a token is closer to the tense or lax class? Desiderata:

* Works in logarithmic space (Mel)
* Speaker-intrinsic, vowel-extrinsic normalization to minimize the effects of physiological differences between speakers (Lobanov)
* Takes into account different variances associated with F1 and F2 (Lobanov)
* Takes into account that F1 and F2 are correlated, especially in the bottom left quadrant of the vowel trapezoid (like Mahalanobis)

Some researchers (e.g., Adank et al. 2004) have proposed the Mel transform as a form of vowel normalization. Unfortunately, this transform does not minimize interspeaker variation due to physiological or anatomical differences, as it does not make use of any speaker-specific statistics.

**data processing steps:**

1. Pull out the subset of white adult (age 18+) speakers
2. Remove function words, using modified list in Selkirk (1984:352-353)
3. Convert all formants from Hz to Mel frequency
4. Z-score Mel formants by speaker (Lobanov normalization)
5. Pull out the subset of primary-stress short-a tokens (AE1)
6. Remove speakers with less than 5 tokens of (æ) or (æh) (because icov problems)
7. Identify speakers with traditional short-a split & only use them for recategorization

**recategorization steps:**

1. Pull out the subset of “æBR” tokens (FAVE-coded as variable words)
2. For each speaker, calculate the Mahalanobis distance between F1, F2 of each of their (æBR) tokens and their (æ) and (æh) means. (Two Mahal dist calculated)
3. Use Mahal dist to calculate whether each word is closer to (æ) or (æh), and code accordingly
4. Tabulate how often each word was coded as (æ) or (æh)
5. Flag the word for recategorization if:
   1. There are at least 5 tokens
   2. The ratio of Mahalanobis dist coding is at least 2:1
6. Incorporate recategorization into shorta.py and run PNC with new coding rules

**Coding Improvements**

What does standard FAVE do?

* Only codes vowels with primary stress
* Tense if followed by word-final or pre-consonantal [M N S TH F] (man, ham; hand, classroom)
* Codes as "variable" if followed by tenser + non-suffix (-ing,-es) vowel
  + Attempt to account for tautosyllabicity without syllabification?
* Variable by stipulation:
  + Lexical exceptions: *math, ran, swam, began, can, family, January, annual, Anne, Joanne, gas, exam, alas, aspirin, Catholic(s), camera*
  + following /l/
  + sC clusters
* Tense by stipulation:
  + *mad, bad, glad* and some derived forms

After removing function words (from Selkirk 1984:352-353):

17% unique words coded as "variable" (447/2,674 short-a words)

* 40% (n=181) of those because AE1 + tensing seg + vowel (*Spanish, planet, damage*)
* 33% (n=149) because AE1 + /l/ (*Italian, pal, alley*)
* 15% (n=65) because AE1 + sC (*basketball, master, faster*)
* 5% (n=23) coded as lexical exceptions (lexically variable) (*family, Catholic, ran)*
* 5% (n=22) resulting from CMU error (*grandmother, Atlantic, advantage*)
* 2% (n=7) were -arry words (*marry, Harry, Larry*)

In terms of tokens, 20% of the short-a data (6,646/33,399) is coded as variable:

* tenser + V, n=1321 or 20%
* lexical exception, n=2252 or 34%
* following /l/, n=1914 or 29%
* following sC, n=716 or 11%
* CMU error, n=240 or 3%
* -arry words, n=203 or 3%

What does SUPER AWESOME KYLARY FAVE do?

* Back-to-basics approach to coding the tense/lax distinction, based on Ferguson. Classes that were coded as “variable” because there's been some change over time or inter-speaker variability have been coded according to Ferguson's original description. These tokens don't need to be excluded from analysis outright as they have been in the past since mixed effects regression models can be used to identify and tease apart any environmental or speaker-level effects / researchers can choose to exclude categories like pre-l that have undergone later change
* Uses stemming to detect when resyllabification has occurred
* Incorporating a syllabification script written by KBG allows for accurate coding of the tautosyllabicity constraint. Replaces hacky FAVE method of checking for tenser + following consonant or vowel.
* Codes all tokens, not just primary stress. FAVE output includes a code for stress, so the investigator can choose to limit the data analyzed or not.
* SC cluster analysis provides coding for tokens previously thrown out as “variable”, identifies true exceptions.
* Not yet handled programmatically: errors caused by multiple pronunciations in CMU dict: schwa-apocope (camera), deletion (Santa, grandmother), weirdness (Africa). The most frequent instances of these currently in the corpus have been coded as exceptions to account for this.

[Figure 1 about here.]

**Analysis of previous variable class categories**

**Tenser + V**

This category no longer exists per se, its members are handled by the syllabification script.

**Lexical exceptions**

These are checked against corpus data and reassigned as necessary.

**Following /l/**

These are coded as lax per orig. desc. In light of Dinkin's work however, researchers using this coding may wish to exclude pre-lateral tokens from analysis.

**Clusters**

Labov 1989 on sC clusters: “These words are normally pronounced with short *a* in an open syllable; otherwise, the stop would be aspirated, which is rarely the case. Syllable structure in the physical sense is not the governing parameter here; rather it is the abstract structure of the word.” [p 24] → see SPE, Vaux, etc on using aspiration as diagnostic of syllable structure. Also Occam's razor → it would be exceptional for short-a in an open syllable to tense in just these words, simplest explanation is that S is in coda.

For s-clusters analysis: 180 speakers from the Philadelphia Neighborhood Corpus (PNC) with clear traditional split short-a system. Yielded a total of 558 tokens of medial sC clusters (where C is [p t k]) following a short-a. For each token we calculated the Mahalanobis distance between it and the speaker's tense and lax short-a means, then coded the token as tense or lax according to which mean it was closer to. Coded tokens were plotted and visually inspected to insure that there were no gross errors in categorization. Finally, these codes were tabulated by word (given in the appendix), using the rule of thumb that words must have at least five tokens to be included (given two equally-probable outcomes, the probability of five instances of the same outcome occurring is .03125), and words are considered truly variable if there are not at least twice as many tokens coded in one class as there are in the other.

from which we may draw the following generalizations:

* SP clusters in this sample promote a lax short-a – only one token of *aspirin* in 28 SP cluster tokens was coded as tense
* ST and SK clusters are majority tense, with a few exceptions: *fantastic, plastic, astronaut, rascal,* and possibly *Alaska. Asteroid* and *elastic* may also fall into this category, but there are too few tokens to be sure. → learned words discussion?
* Thus it would appear that in ST and SK clusters, Philadelphians are analyzing the S as a coda consonant, and tensing short-a due to the tautosyllabic S. More data on SP clusters is needed, but the results here suggest that they behave differently, with S analyzed in the onset of the following syllable, thus not causing short-a tensing.

So what went into our short-a coding is: S is syllabified in coda, so normally coded as tense, with the lexical exceptions *aspect*, *aspirin*, *rascal,* which are lax, and *plastic, Alaska, fantastic* which are variable.

**CMU error**

Impossible to catch all of these without manually fixing the dictionary; most frequent affected words currently in the corpus have been hard-coded as exceptions in the relevant category.

**-arry words**

Coded as lax per Labov.

**Results**

**Conclusions**

**Appendix**

List of short-a words included in cluster analysis and Mahalanobis distance categorization:

*Threshold for classification: at least 5 tokens & a 2:1 difference*

**Word (lax) ae aeh**

ASPECT[S] 8 1

ASPIRIN 16 1

ASTERISK 4 1

ASTEROID[‘S/S] 5 0

**Word (tense) ae aeh** ALASKA[N] 1 6

ASK[ED/ING] 5 178

BASKET[S] 0 18

BASKETBALL 2 77

CASKET 0 5

DISASTER 0 7

DRASTICALLY 1 5

FAST[ER/EST] 1 13

LAST[ED/ING] 3 14

MASTER[‘S/S] 7 29

NASTY 2 11

PASTOR[‘S/S] 2 17

PLASTER 1 4

PLASTIC 1 9

TASKER 0 25

**Word (variable?) ae aeh** FANTASTIC[ALLY] 9 5

**Figures**

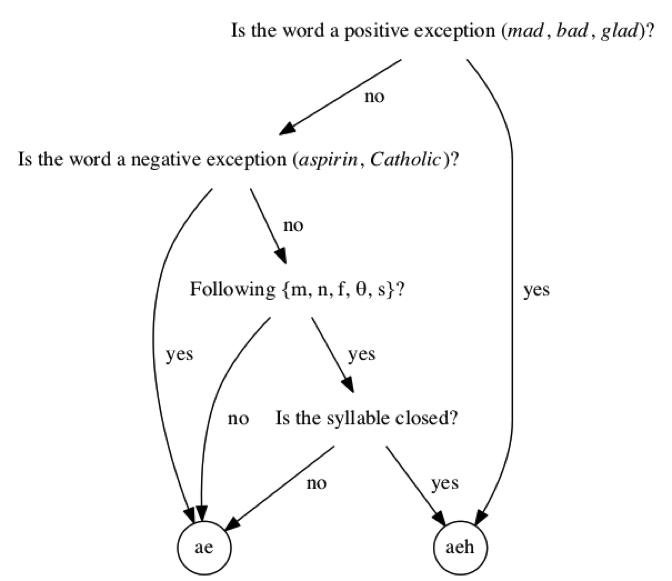


Figure 1: Proposed scheme for automated short-a coding.