Awesome short-a funtimes

Hilary Prichard (University of Pennsylvania)

Kyle Gorman (OHSU)

**Introduction**

**Short-a in Philadelphia**

Ferguson 1972 gives the original complete description 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 informant responses. Ferguson first defines the phonetic quality of the two phonemes which make up the split system, noting that (æ) is low, front, unrounded, monophthongal, and 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) *æ → æh / \_* {m n f θ s}{*C* #} # insert a picture of the proper rule??

Ferguson continues by examining the 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” *Afghan, aft, asp, crass, daft, damsel, gaff(e), Gath, hasp, lass, lath, Rasputin, tam, Tass,* 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 morphologically-conditioned exceptions, such as the “preservation of stem identity” leading to tensing in words with an *-ing, -es, -ed,* and possibly *-er* suffix, like *passing, classes.* Finally there are exceptions to tensing which result from shortening (*math, exam*), elision (*family, camera, Catholic*), or the special case of short-a in an initial stressed syllable followed by *-sp, -sf,* or *-sb* (*aspirin, asphalt, asbestos*).

Labov 1989 and 1994 build on Ferguson's description, formalizing additional rules, and investigating possible exceptions using the Philadelphia LVC data. Labov introduces the idea of tautosyllabicity as necessary for tensing; where Ferguson contrasted tensing consonants followed by another consonant (tense) with those followed by a vowel (lax), Labov makes this description more precise by stating that tensing only occurs in closed syllables (1994: 430). He adds that Level 2 inflectional suffixes do not cause laxing, while Level 1 derivational suffixes show considerable variation (with the unfortunate examples *plastic* and *Lassie).* Labov also finds variation in words with ST clusters (*master, plaster*), and possible lexical diffusion of NV and LV clusters into the tense class (*plant, 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.

**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" (476/2,853 short-a words) # three examples of each

* 40% (n=189) of those because AE1 + tensing seg + vowel (*Spanish, planet, damage*)
* 34% (n=163) because AE1 + /l/ (*Italian, pal, alley*)
* 15% (n=71) because AE1 + sC (*basketball, master, faster*)
* 5% (n=23) coded as lexical exceptions (lexically variable) (*family, Catholic, ran)*
* 5% (n=22) resulting from CMU t/d deletion after /n/ (*grandmother, Atlantic, advantage*)
* 2% (n=8) were -arry words (*marry, Harry, Larry*)

In terms of tokens, 13% of the short-a data (6,966/54,125) is coded as variable:

* tenser + V, n=1314 or 19%
* lexical exception, n=2528 or 36%
* following /l/, n=1884 or 27%
* following sC, n=752 or 11%
* CMU error, n=303 or 4%
* -arry words, n=185 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.

**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:

#odds ratio of four (one is twice as likely as the other) is the threshold for inclusion

Word (exceptions) æh æ

**ASP**IRIN(S) 1 13

**ASP**ECT(S) 0 10

FANT**AST**IC 4 9

PL**AST**IC(S) 4 4

AL**ASK**[A/AN] 4 3

R**ASC**AL 0 4

Word (tensed) æh æ

M**AST**ER[S/'S] 30 11

P**AST**OR[S/'S] 17 2

F**AST**[ER/EST] 16 2

L**AST**ED 11 5

N**AST**Y 11 2

PL**AST**ER(ED) 7 1

DIS**AST**ER 6 1

Word (tensed) æh æ

DR**AST**ICALLY 5 1

**ASK**[ED/ING] 174 15

B**ASK**ETBALL 75 3

T**ASK**ER 28 1

B**ASK**ET(S) 23 0

C**ASK**ET(S) 6 0

Word æh æ

**ASP**IRIN(S) 1 13

**ASP**ECT(S) 0 10

C**ASP**ER 0 2

**ASP**EN 0 1

**ASP**IRATIONS 0 1

M**AST**ER[S/'S] 30 11

P**AST**OR[S/'S] 17 2

F**AST**[ER/EST] 16 2

L**AST**ED 11 5

FANT**AST**IC 4 9

N**AST**Y 11 2

PL**AST**ER(ED) 7 1

PL**AST**IC(S) 4 4

DIS**AST**ER 6 1

DR**AST**ICALLY 5 1

**AST**RO[/NAUT] 0 4

Word æh æ

**AST**ERISK(S) 1 2

BL**AST**ED 3 0

M**AST**ERMAN 3 0

SARC**AST**IC(ALLY) 2 1

**AST**EROID(S) 0 2

B**AST**ARD 2 0

C**AST**ING 1 1

C**AST**OR 2 0

EL**AST**IC 0 2

LANC**AST**ER 1 1

BROADC**AST**ERS 1 0

C**AST**ER 1 0

CH**AST**ISED 0 1

DIS**AST**ROUS 0 1

ENTHUSI**AST**IC 1 0

F**AST**ING 1 0

P**AST**ORAL 1 0

Word æh æ

P**AST**URE 1 0

SANDBL**AST**ED 0 1

SCHOL**AST**IC 0 1

**ASK**[ED/ING] 174 15

B**ASK**ETBALL 75 3

T**ASK**ER 28 1

B**ASK**ET(S) 23 0

AL**ASK**[A/AN] 4 3

C**ASK**ET(S) 6 0

R**ASC**AL 0 4

G**ASK**ELL 2 0

M**ASC**OT 1 1

FI**ASC**O'S 1 0

M**ASC**ULINE 1 0

M**ASK**IN' 1 0