2017-07-01

A Latent Class Analysis of American English Dialects

Stephanie Nicole Hedges
Brigham Young University

A Latent Class Analysis of American English Dialects

Stephanie Nicole Hedges

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

David S. Eddington, Chair
Dirk Allen Elzinga
Deryle W. Lonsdale

Department of Linguistics and English Language

Brigham Young University

ABSTRACT

A Latent Class Analysis of American English Dialects

Stephanie Nicole Hedges
Department of Linguistics and English Language, BYU
Master of Arts

Research on the dialects of English spoken within the United States shows variation regarding lexical, morphological, syntactic, and phonological features. Previous research has tended to focus on one linguistic variable at a time with variation. To incorporate multiple variables in the same analysis, this thesis uses a latent class analysis to perform a cluster analysis on results from the Harvard Dialect Survey (2003) in order to investigate what phonetic variables from the Harvard Dialect Survey are most closely associated with each dialect. This thesis also looks at how closely the latent class analysis results correspond to the Atlas of North America (Labov, Ash & Boberg, 2005b) and how well the results correspond to Joshua Katz's heat maps (Business Insider, 2013; Byrne, 2013; Huffington Post, 2013; The Atlantic, 2013).

The results from the Harvard Dialect Survey generally parallel the findings of the Linguistic Atlas of North American English, providing support for six basic dialects of American English. The variables with the highest probability of occurring in the North dialect are 'pajamas: /æ/', 'coupon: /ju:/', 'Monday, Friday: /e:/' 'Florida: /ɔ/', and 'caramel: 2 syllables'. For the South dialect, the top variables are 'handkerchief: /ɪ/', 'lawyer: /ɒ/', 'pajamas: /ɑ/', and 'poem' as 2 syllables. The top variables in the West dialect include 'pajamas: /ɑ/', 'Florida: /ɔ/', 'Monday, Friday: /e:/', 'handkerchief: /ɪ/', and 'lawyer: /ɔj/'. For the New England dialect, they are 'Monday, Friday: /e:/', 'route: /ru:t/', 'caramel: 3 syllables', 'mayonnaise: /ejɑ/', and 'lawyer: /ɔj/'. The top variables for the Midland dialect are 'pajamas: /æ/', 'coupon: /u:/', 'Monday, Friday: /e:/', 'Florida: /ɔ/', and 'lawyer: /ɔj/' and for New York City and the Mid-Atlantic States, they are 'handkerchief: /ɪ/', 'Monday, Friday: /e:/', 'pajamas: /ɑ/', 'been: /ɪ/', 'route: /ru:t/', 'lawyer: /ɔj/', and 'coupon: /u:/'. One major discrepancy between the results from the latent class analysis and the linguistic atlas is the region of the low back merger. In the latent class analysis, the North dialect has a low probability of the 'cot/caught' low back vowel distinction, whereas the linguistic atlas found this to be a salent variable of the North dialect. In conclusion, these results show that the latent class analysis corresponds with current research, as well as adding additional information with multiple variables.

Keywords: American English dialects, latent class analysis, dialect variation

ACKNOWLEDGEMENTS

I would like to thank my committee chair David Eddington, as well as my committee members Dirk Elzinga and Deryle Lonsdale for their continuous support and direction.

I would also like to offer sincere thanks to Chongming Yang, who aided understanding and completing of the statistical analyses.

Finally, I would also like to thank Bert Vaux and Scott Golder for generously allowing me access to their data, without which this thesis would not have been possible.

TABLE OF CONTENTS

INTRODUCTION

Over time, human language and communication has changed and likely will continue to do so. This phenomenon is known as language shift. One aspect of language shift includes variations of pronunciation and speech. The reasons for such variation include a dynamic interplay of many social aspect including politics, social identity, and social change (Coloma, 2012; Edwards, 2009; Labov, 1963; Lakoff, 1990; Skendi, 1975).

In regards to language, Charles Ferguson (1994) noted that "A group that operates regularly in a society as a functional element will tend to develop identifying markers of language structure and language use". According to this model, dialects are a result of geographical location, economic position, and the historical era. Several studies including the Atlas of North American English (Labov, Ash & Boberg, 2005b) works mostly with dialects defined based on physical location and not dialects resulting from religious, economic, or historical factors, even though these factors may influence the dialect of physical locations.

Research on the dialects of English spoken within the United States shows variation regarding lexical, morphological, syntactic, and phonological features. For example, Metcalf (2000) identified a lexical variation in "seesaw" used in Southern and Midland dialects and "dandle" used in Rhode Island, while "teeter totter" is used throughout the United States. Furthermore, phonological dialect variation includes the occurrence of pronouncing /t/ as a glottal stop (Eddington & Channer, 2010) as well as variation in vowel formant frequency (Hagiwara, 1997). Additionally, Grieve (2012) found syntactic variation involving the placement of adverbs between the Northeast, the Southeast, and the South Central states in the United States.

However, many previous studies focus only on how dialects are similar or dissimilar in regards to just one or two linguistic features. Yet dialects are complex, and multiple linguistic features combine to make a specific dialect. A dialect can be both similar and dissimilar in relation to other dialects depending on the investigated variables. For example, using a glottal stop in place of a /t/ is more common in the Western dialect than a non-western dialect of the United States (Eddington & Channer, 2010); yet both the Western dialect and the Southern dialect use the word "milkshake/shake" for a drink made with milk and ice cream (Vaux & Golder, 2003). Because of multiple features combining to create a dialect, a multivariate statistical analysis should be used when investigating regional variation in American English. Multivariate analysis is a tool that analyzes data with several variables. These techniques have arisen with the development of computers that are capable of computing large amounts of data (Abdi, 2003). Multivariate analysis applies to dialectology as it is able to take into account each linguistic variable to establish dialect boundaries.

While there are many studies in dialectology using multivariate analyses (Wieling & Nerbonne, 2015), this thesis uses a latent class analysis to perform a cluster analysis on results from the Harvard Dialect Survey (2003), a dataset from a survey eliciting for phonetic variation within the United States. The results from this analysis allow me to investigate the following two questions:

1. What phonetic variables from the Harvard Dialect Survey are most closely associated with each dialect?
2. How closely do the results from The Harvard Dialect Survey correspond to the Atlas of North America (Labov, Ash & Boberg, 2005b) and specifically the dialect regions established in the atlas, and also how well the results correspond to Joshua Katz's

heat maps (Business Insider, 2013; Byrne, 2013; Huffington Post, 2013; The Atlantic, 2013) produced using the same data from the Harvard Dialect Survey, but not separated into dialect regions?

LITERATURE REVIEW

The Atlas of North American English

The Atlas of North American English (Labov, Ash & Boberg, 2005b) is the first comprehensive linguistic atlas of English spoken in North America. Its contents build on the results of past studies as well as new phonetic and perceptual data. The new data were collected between 1992 and 1999 by the Telsur phone survey with a sample of over 50,000 participants. Most of the participants were from larger, urbanized cities; however, a small amount of smaller populated areas is also included in order to best represent the language of North America.

The criteria the linguistic atlas uses for dividing North America into dialect regions includes vowel position and sound changes such as mergers, splits, and chain shifts. The results of the dialect regions can be seen in Figure1.

Figure 1: *The Linguistic Atlas of North American English* dialect regions (Labov, Ash & Boberg, 2005b).

As the map shows, the linguistic atlas identified several dialect regions based on geography in the United States. However, the linguistic atlas identifies six major dialect regions:

the North, New England, New York City and the Mid-Atlantic States, the South, the Midland, and the West (Labov, Ash & Boberg, 2005a). Like Labov, Ash & and Boberg, I will include the sub regions identified in the linguistic atlas in Figure 1 into these six regional dialects in order to simplify the latent class analysis (Labov, Ash & Boberg, 2005a). The North dialect will represent the dialect spoken in the Inland North, as well as the North Central region. The New England dialect will represent both dialects from Eastern New England and Western New England. New York City and the Mid-Atlantic States dialect includes the New York City region and the Mid-Atlantic States region. The South dialect includes the Texas South, Florida, Charleston, and the Inland South. The Midland dialect also includes the St. Louis Corridor and Western Pennsylvania. And the West is its own dialect region.

The North

The Linguistic Atlas of North American English found that the phonetic variable that distinguished the North dialect from other dialects in the United States is the presence of the Northern Cities Shift. The Northern Cities Shift is a chain shift of lax vowels in American English. This shift is initiated by the raising and fronting of /æ/. This vowel's movement allows /ɑ/ to become fronted, followed by the lowering of /ɔ/, and the lowering and backing of /ɛ/, and finally the backing of /ʌ/ (Labov, Ash & Boberg, 2005a; Labov, Ash & Boberg, 2005b).

The North dialect also is distinguished by the absence of the low back vowel merger of the vowels /ɑ/ and /ɔ/ to be pronounced as /ɑ/. These two vowels are distinct in the North dialect so the pronunciations of 'dawn' and 'Don' are pronounced with distinct vowels (Labov, Ash & Boberg, 2005b).

New England

When describing the New England dialect, the linguistic atlas divides the dialect region into four quadrants. To make these quadrants, a horizontal line representing the low back vowel (/ɑ/ and /ɔ/) merger divides the North from the South, where Northern New England has the low back vowel merger and Southern New England distinguishes between these two low back vowels. R-vocalization represents the vertical line where Eastern New England has an r-vocalization where a /r/ is pronounced as a vowel and Western New England being an r-full dialect. So, to describe each quadrant region, Northeastern New England (including Boston and the surrounding area) has r-vocalization in the dialect with the low back merger as well. The linguistic atlas also found Northeastern New England to front the /ɑ/ vowel in 'father', 'pajama', 'aunt', etc. Southeastern New England also has r-vocalization, but does not have the low back merger of the vowels /ɑ/ and /ɔ/. Similar to the Northeast, Northwestern New England also is distinguished by the low back merger; however, it is an r-full dialect. The fourth quadrant, Southwestern New England, does not have the low back merger and is an r-full dialect (Labov, Ash & Boberg, 2005b).

Western New England is also characterized by speech whose difference between F2 of the mid vowels /e/ and /o/ is less than 375 Hz. In this situation, /e/ is backed whereas /o/ is fronted. This vowel characteristic is also found in the Northern Cities Shift. However, while the Northern Cities Shift's mid vowel movement is driven by the raising of /æ/, the vowel movement in Western New England is not driven by an encroaching vowel (Labov, Ash & Boberg, 2005b).

New York City and the Mid-Atlantic States

The linguistic atlas groups New York City and the Mid-Atlantic States regions together because of two shared linguistic features: the raising of /ɔ/ to a mid-high position (and thus resisting the low back merger) and a split short-a /æ/. In both of these regions, the short /æ/ splits into either lax /æ/ or tense /æə/ (Labov, Ash & Boberg, 2005b).

However, the dialect of New York City behaves differently than the dialect of the Mid-Atlantic States in the nature of the short-a split and the vocalization of /ɹ/. Several studies have identified multiple contexts where lax /æ/ becomes tense /æə/ such as in closed syllables before nasals and voiceless fricatives and before /d/, while the short-a is lax /æ/ in auxiliaries and irregular verbs with nasal codas (i.e. 'ran) (Banuazizi & Lipson, 1998; Ferguson, 1975; Labov, 1989; Labov, Ash & Boberg, 2005b; Roberts, 1993; Roberts & Labov, 1995). New York City is typically r-less while the Mid-Atlantic States generally pronounce /ɹ/ (Labov, Ash & Boberg, 2005b).

The South

The linguistic atlas characterizes speech in the South as combining several phonetic variables. For instance, the Southern dialect is rhotic in syllable final positions. Also in syllable final positions with the suffix '-ing', the nasal takes an alveolar place of articulation instead of a velar one. One of the most noticeable characteristic of the Southern dialect is the relatively high use of glides. For example, /æ/ before sibilants and nasals is often upglided to /æj/. Furthermore, the sound /uw/ often has the glide /j/ added to the front becoming /juw/ following coronals in the same syllable. For example, 'tune' would take the pronunciation of /tjuwn/ (Labov, Ash & Boberg, 2005b).

Other characteristics of the South that the linguistic atlas mentions include the fronting of back vowels. The vowels /u/, /uw/, and /ow/ are all fronted. The diphthong /aʊ/ is also fronted to /æw/ as in 'out' and 'mountain'. Similarly, the back vowel /ɔ/ is upglided to /ɔw/ as in 'caught' and 'law' (Labov, Ash & Boberg, 2005b).

The linguistic atlas also found the South to also be distinguished by the presence of what is known as the Southern Vowel Shift. This shift begins with the diphthong /æj/ to the monophthong /æ/. Also, the nucleus of the diphthong /ej/ is lowered. This allows for /i/, /ɛ/, and /æ/ to become raised and fronted as well as having an inglide. This creates the effect of what is known as the stereotypical Southern drawl (Labov, Ash & Boberg, 2005a; Labov, Ash & Boberg, 2005b).

Several sound distinctions are characteristic of the Southern dialect. The linguistic atlas found that the South distinguishes between /hw/ and /w/, the most famous example being 'which' and 'witch'. Furthermore, the distinction between the vowels in 'marry' and 'merry' are maintained as /æ/ and /e/ respectively. This region also maintains the distinction of the low back vowels found in 'cot/caught' and 'Don/dawn' (Labov, Ash & Boberg, 2005b).

The linguistic atlas also found several vowel mergers occurring in the Southern dialect. For instance, the South has what is commonly known as the "pin/pen" merger where the vowels /ɪ/ and /ɛ/ are merged before nasals. The vowels /u/ and /ʊ/ are merged before /l/, causing a similar pronunciation of the words 'pull' and 'pool' and also 'full' and 'fool'. The vowels /ɛ/ and /ej/ are also merged before /e/, as well as /i/ and /ɪ/ as in the words 'sell' and 'sail' as well as in 'feel' and 'fill' (Labov, Ash & Boberg, 2005b).

The Midland

The linguistic atlas characterizes the Midland as a dialect region where the low back merger is transitional, meaning that it is merged in some contexts (largely geographical), but not completely merged in other contexts. However, the Midland is also a dialect region that is large and has very distinct dialects occurring in individual cities such as Pittsburgh and St. Louis. Because of this, the characteristics of the Midland mentioned are very broad and cannot necessarily be assumed for the entire region. In Pittsburgh, for example, the low back merger is complete and not in transition, as it is in the majority of the dialect region (Labov, Ash & Boberg, 2005b).

The Midland is also characterized by the fronting of the diphthongs /aw/ and /ow/ as well as /ʌ/. Also, it is marked by glide deletion before sonorant consonants. However, this phonetic characteristic is also in transition in the Midland where the northern region has less glide deletion and the southern region has a greater percentage of glide deletion (Labov, Ash & Boberg, 2005b).

The West

According to the linguistic atlas, the most salient linguistic characteristic of the West dialect is the presence of the low back merger. Another characteristic that the linguistic atlas mentions is that the vowel /uw/ is fronted; however, the parallel vowel /ow/ is not. This is different in other American dialects where both of the vowels are fronted together. The linguistic atlas also found the West dialect to be a "dialect area with low homogeneity and moderately low consistency" meaning that the dialect within the West varies considerably between cities throughout this dialect region. (Labov, Ash & Boberg, 2005b).

The Harvard Dialect Survey

Like the Atlas of North American English, the Harvard Dialect Survey created by Bert Vaux and Scott A. Golder also elicited for differences in English dialects spoken across the United States. The survey was distributed online and completed in 2003 (Vaux & Golder, 2003).

The entire survey is compiled of 122 questions regarding phonetic, lexical, syntactic, and morphological differences in English in the United States. The questions are multiple-choice with a write-in option if the participant's pronunciation of the elicited feature was not already a choice. The questions use rhyming words in order for the participants to best pick the option with their true pronunciation. For example, Question 7 from the survey elicits for the pronunciation of the first vowel in '**c<u>ou</u>**pon' with the options "(a) with /u:/ as in "coop" ("coopon"); (b) with /ju:/ as in "cute" ("cyoopon"); or (c) other" (Vaux & Golder, 2003).

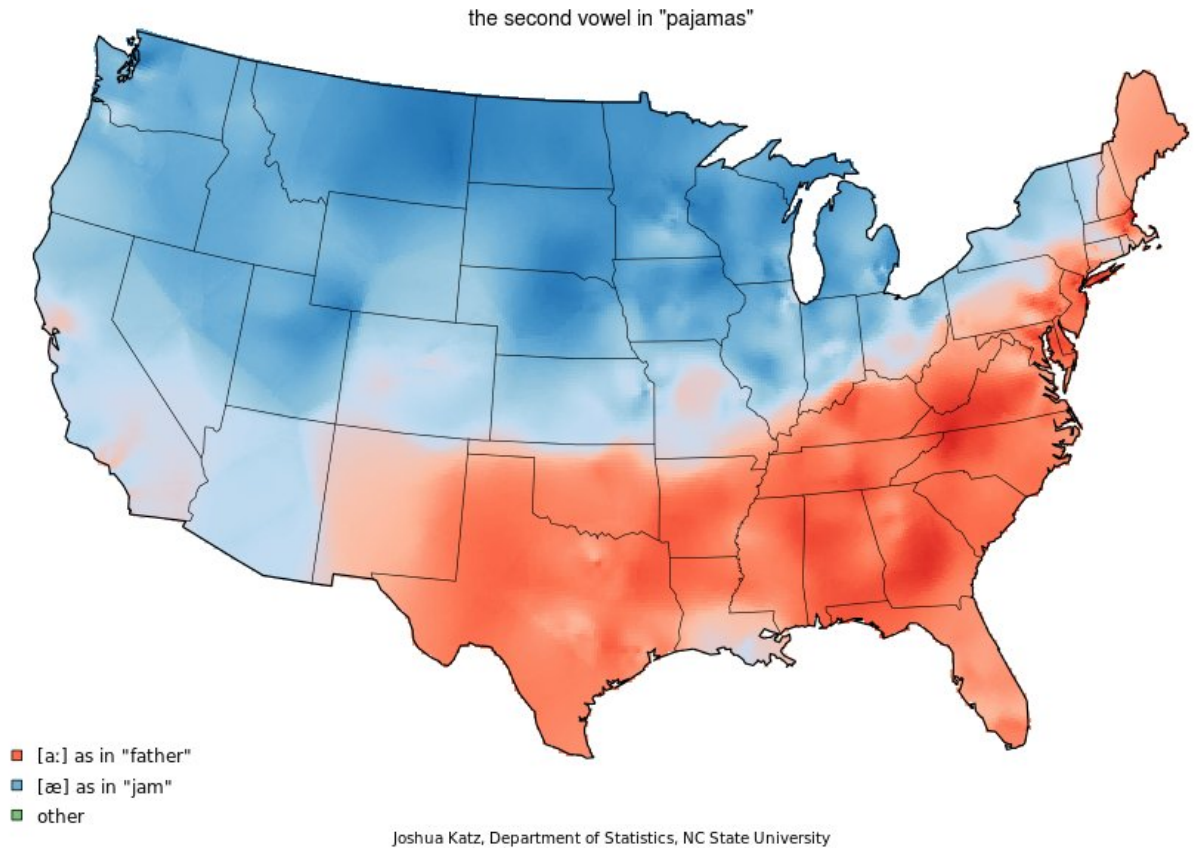
Each state was represented by between 68 (Hawaii) and 2773 (California) participants. The total number of participants was 30,788. The participants were between ages of 13 and 70+ (Vaux & Golder, 2003).

Joshua Katz's Heat Maps

The Harvard Dialect Survey gained popularity among Americans in 2013 when Joshua Katz, then a doctorate student of statistics at North Carolina State University, generated heat maps for the data, excluding Alaska and Hawaii. The heat maps allowed for better visualization of the data as they took population density into account (Katz, 2013). Katz's Heat Maps were first published in North Carolina State University's research journal *The Abstract* where they were then picked up by multiple news agencies across the United States such as *Business Insider*, *The Atlantic*, the *New York Times*, the *Huffington Post*, and *New York Daily News.*

All of the heat maps generated by Katz are not available for public viewing. However, several of the maps can be viewed on the various news sites. The available heat maps show general trends in the six dialect regions defined by the linguistic atlas (Business Insider, 2013; Byrne, 2013; Huffington Post, 2013; The Atlantic, 2013;). An example of a heat map can be seen in Figure 1, and the complete available heat maps are in the Appendix.

FIGURE 1: Heat map of the variable 'pajamas'.



The North

Joshua Katz's heat maps of the dataset show a lowering and backing of /e/ into the vowel /ɛ/ as seen in the pronunciation of 'been' in the North dialect (Business Insider, 2013). Also, they show the resistance to the low back merger in the majority of the North dialect, especially in Wisconsin, Michigan, and Western New York, distinguishing between the two vowels in

‘cot/caught’ (Byrne, 2013). The low front vowel is also relatively front and raised as seen in the pronunciation of ‘pajamas’ and ‘aunt’ with /æ/ as opposed to /a:/ or /ɑ/ (Business Insider, 2013).

New England

The heat maps show the New England dialect as pronouncing ‘aunt’ and the second vowel in ‘pajamas’ with /a/ and the presence of the low back merger except for Connecticut and Rhode Island (Business Insider, 2013). Also, according to the maps, the New England dialect pronounces ‘lawyer’ with the /ɔj/ vowel rather than with /ɑ/. The pronunciation of ‘been’ seems to be mixed between /ɪ/ and /ɛ/, leaning more towards /ɪ/, especially in Boston and the surrounding area (Business Insider, 2013).

The South

Katz’s heat maps show /ɔj/ pronounced as the monophthong /ɑ/ as in ‘lawyer’ as well as the diphthong /ejə/ pronounced as a monophthong /æ/ in ‘mayonnaise’ in the South dialect (Business Insider, 2013; The Atlantic, 2013). Also, like the linguistic atlas, the heat maps found a strong presence of the low back merger in the South dialect region (Byrne, 2013).

New York City and the Mid-Atlantic States

The heat maps show New York City and the Mid-Atlantic States dialect to resist the low back merger (Byrne, 2013). The maps also show that this region is unique in its pronunciation of the vowel before the /ɹ/ in ‘syrup’ with the vowel /i/ or /ɪ/ where the rest of the dialects have /ə/ as the likely pronunciation (The Atlantic, 2013). Also ‘aunt’ is shown to have the /æ/ pronunciation and ‘pajamas’ the /ɑ/ pronunciation (Business Insider, 2013).

The Midland

The heat maps show the Midland dialect to share several of the phonetic variables with the South dialect and others with the North dialect. For example, the Midland dialect is more similar to the South dialect in its pronunciation of 'been' with /ɪ/ and 'mayonnaise' with /æ/ (Business Insider, 2013; The Atlantic, 2013). However, its pronunciation of 'pajamas' with /ɑ/, 'lawyer' with /ɔj/, and 'caramel' as 2 syllables are more similar to the North dialect's pronunciation (Business Insider, 2013).

The West

The heat maps show the West dialect to have merge the low back vowels /ɑ/ and /ɔ/in 'cot/caught', similar to the findings in the linguistic atlas (Byrne, 2013). It also shows a strong pronunciation of the vowel in 'aunt' and the second vowel in 'pajamas' to be pronounced with /æ/ (Business Insider, 2013).

Multivariate Approaches in Dialectology

Biber (1985) used multidimensional analysis in linguistics with his research of register variation. Since then, many linguists have used multidimensional analysis to study language feature co-occurrences. Hyvönen et al. (2007) performed a multivariate analysis on a comprehensive dictionary of Finnish regional dialects to better understand the variation of dialects based strictly on lexical items. Additionally, in 2009, Xiao applied multidimensional analysis to synchronic data of world-wide English variation using the International Corpus of English (ICE).

The factor analysis used in multidimensional analysis reduces the data to representative features, or variables, in terms of factor loadings for each of the dimensions, or underlying

factors, it creates from the data. Factor loadings are numbers between -1 and 1 that show how well a linguistic feature is represented in a dimension. The closer the loadings are to -1 or 1, the stronger effect the feature has in the dimension. For this thesis, a factor analysis would group phonetic features that statistically occur together. Features in this thesis are the particular phonetic variables of interest. However, because factor analysis ignores similarities between individual cases, a better suited statistical method is a cluster analysis, as it takes into account a person's unique phonetic pattern that aligns with a particular dialect. This combination of all the phonetic variables (or features) for one person is known as a case.

A cluster analysis reduces the data into statistically associated cases and measures the probability of each phonetic feature occurring in each cluster (representative case). By reducing the data into representative cases, I can investigate dialects using a variety of phonetic features simultaneously (Conduct and Interpret a Cluster Analysis, 2017).

There are several types of cluster analyses. Bacher (2004) evaluated a common cluster analysis called the TwoStep cluster analysis in terms of type of data and performance of analysis. This evaluation gave evidence that the cluster analysis performed well when the data were continuous. However, if the data were not continuous, the results were unsatisfactory, as the differences between categorical variables were given greater weight, skewing the results. Bacher suggested a latent class model instead of a TwoStep cluster analysis to reduce data to a representative case with categorical data. Because of this finding, this thesis will use a latent class model instead of the TwoStep cluster analysis to group the dataset into statistically representative cases and produce numeric data indicating the probability of phonetic features occurring in each representative case.

By using the latent class model to analyze the data from the Harvard Dialect Survey, I will be able to separate out six American English dialects from the data and see the probability of each linguistic feature occurring in each dialect in order to answer my first research question of which phonetic variables are most associated with each dialect. The probabilities of each linguistic feature occurring in each dialect produced by this model will also allow me to investigate my second research question by using the data to compare with the findings from the Atlas of North American English as well as Katz's heat maps. Specifically, I will determine how the clusters that the analysis groups together match the features the survey shows are dialectal features in each dialect region from the heat maps. I will also be able to compare how closely each dialect is to another using a correlation analysis.

METHODOLOGY

Participants

This study makes use of the online survey results from the Harvard Dialect Survey that Bert Vaux finished conducting in 2003 (Vaux & Golder, 2003). Each state in the United States had between 70 and 2773 participants. The survey contained 122 multiple-choice questions collecting data on lexical, phonetic, and syntactic variations of English spoken across the United States. An example of a question containing phonetic data was Q20: How do you pronounce the second vowel in "pajamas"? (/æ/ as in "jam", /ɑ/ as in "father", or other). This current study will only use the 55 questions asking for the phonetic variation that occurs throughout the United States. Each phonetic question in the survey had between 10632 and 11713 respondents. A copy of the phonetic questions used from the survey can be found in the appendix of this thesis.

Statistics

Latent Class Model

To classify the dialects of English in the United States, I performed a latent class analysis using the software Mplus 7.4 (Muthén & Muthén, 2015) on the data. A latent class analysis works similarly to a factor analysis in that it reduces the data and accounts for how data points interact with each other. However, it differs from a factor analysis in terms of how it reduces the data. Instead of reducing the data into representative features as in a factor analysis, a latent class analysis reduces the data to representative cases, or clusters, and measures each variable in terms of its probability of occurring in each cluster. Or in other words, this analysis will show the probability that a phonetic feature will occur in each cluster, i.e., dialect. By reducing the data to a representative case, I am able to investigate American English dialects as a combination of linguistic features. I make the assumption that the clusters can be interpreted as separate dialects.

Preprocessing Data

I recoded the questions in order to make the answers binary to make the analysis more straightforward. I achieved this by breaking an individual question from the survey into multiple questions where each answer choice is a separate question with 0 representing when the feature does not exist and 1 representing when it does exist. For example, Question 108 was recoded into three questions.

The original question is:

108. What vowel do you use in b*a*g?

a. /æ/ as in "sat"
b. /ɛ/ as in "set"
c. /e:/ as in "say"

The recoded, binary questions are:

Question 108a: /æ/, Choices (a) = 1, all other choices = 0
Question 108b: /ɛ/, Choices (b) = 1, all other choices = 0.
Question 108c: /e:/, Choices (c) = 1, all other choices = 0.

I used Mplus to run the latent class analysis. Mplus measures the uncertainty of the model by its relative entropy, a number between 0 and 1 where the values closer to 1 indicate a higher certainty of the data fitting the model (Kupzyk, 2011). As I ran the analysis using five, six, seven, and eight clusters, and I found that the more clusters there were, the higher the relative entropy of the analysis, or the better the variables fit into clusters. This is because the relative entropy will continue to increase as fewer variables are expected to fit into each cluster. I decided to choose six clusters because the Atlas of North American English divides the United States into six major dialect regions: New England, New York City and the Mid-Atlantic States, the North, the Midland, the South, and the West. Additionally, I changed all transcriptions into IPA in order to make the results and interpretations consistent.

After running the first analysis, I excluded questions that showed little phonetic variation between the cases. I then ran a second latent class analysis for six and seven clusters with the remaining phonetic questions. Again, I removed questions that showed little phonetic variation in order to select the variables that varied the most throughout the dialects.

Labeling of Clusters

To label the clusters, I will align the data from the latent class analysis with the heat maps, as well as with the findings from the Linguistic Atlas of North America.

Correlation Analysis

To compare the similarities and differences within the different clusters from the latent class analysis with each other, I will run a correlation analysis on SPSS (IBM, 2015). Using the phonetic feature's probabilities of occurring in each of the six clusters, i.e., dialects, I will generate a correlation table showing the correlation of phonetic features between dialect clusters. This will show how closely each dialect is related to the others based on the phonetic features from the survey.

RESULTS

Labeling the Clusters

To best interpret the clusters, I compared the probabilities obtained from the latent class analysis showing the probability of each phonetic variable occurring in each dialect cluster with the heat maps produced by Joshua Katz by visual inspection, as well as with the Atlas of North American English. Both of these sources aided in my decision of what to label the clusters.

I hypothesize that the six clusters from the latent class analysis should align with the six major dialect regions within the United States because the latent class analysis separated the data into the best fit for six clusters and the linguistic atlas has six major dialect regions because each dialect region behaved more similarly than with other dialects and different from the other dialect regions. Joshua Katz's heat maps provided a visual representation of the data found by the Harvard Dialect Survey of Bert Vaux. Unfortunately, Katz's original maps are no longer available for public use. I could find access to only some of the heat maps through news/magazine articles such as the *Huffington Post* and *Business Insider*. I could only find fourteen out of the twenty five variables that I am using in the latent class analysis. The variables from the survey that I have heat maps for include "aunt", "been", "Bowie knife", "caramel", "crayon", "cot/caught", "coupon", "lawyer", "mayonnaise", "pajamas", "pecan", "route", and "syrup". This left the variables "cauliflower", "Craig", "creek", "Florida", "flourish", "handkerchief", "miracle", "Monday, Friday, etc.", "poem", "really", and "realtor" without a visual comparison.

With these heat maps, I noticed the areas where specific pronunciations for a linguistic variables occurred. I then counted the number of times the cluster's probability matched up with each dialect. For example, with the variable "lawyer", the pronunciation /ɔj/ was relatively high

in Cluster I (p=0.841), III (p=0.870), IV (p=0.664), V (p=0.879), VI (p=0.880), but lower in Cluster II (p=0.141). To match them up, I visually inspected the heat maps. If I noticed that the region had more of one color then it matched if the feature's probability of the corresponding main color had a higher probability than other possible pronunciations of the variable. For example, in the case of the variable "lawyer", the heat map shows that the pronunciation with /ɔj/ occurred in the West, the North, the Mid-Atlantic States, New England, and to some extent the Midland dialect, whereas the pronunciation including /ɑ/ occurred in the South and a little in the Midland. Each time the variable's probability correctly matched up with a dialect based on the trends from the heat maps, I gave that cluster a point for that dialect. Ideally, each dialect would match up with a separate cluster (Table 1). So, for the example of "lawyer", Cluster II matched up with The South, and the other clusters matched up with the remaining dialect labels.

Table 1: Linguistic variables in dialect regions from the heat map compared with the linguistic variables from latent class analysis.

	I	**II**	**III**	**IV**	**V**	**VI**
Midland	22*	13	20	11	20	7
South	12	22*	12	8	10	12
West	21	11	21	12	23*	11
New England	12	9	15	20*	13	19
North	19	14	15	12	19	12
NYC & Mid-Atlantic States	8	14	14	17	12	24*

The asterisks (*) represent where the largest count in the row was also the largest count in the column.
The Roman numeral headings represent the clusters produced from the latent class analysis.

Where the column and row shared the same cell with the highest number, the dialect matched up with the cluster, according to the heat maps. This occurred five times: the Midland

dialect with cluster I, the South dialect with cluster II, the New England dialect with cluster IV, the West dialect with cluster V, and the New York City and the Mid-Atlantic States dialect with cluster VI. After consulting the linguistic atlas, I decided to label cluster II as the South, cluster IV as New England, and cluster VI as New York City and the Mid-Atlantic States. However, findings from the linguistic atlas suggested not to label cluster I as the Midland and cluster V as the West. With cluster I, cluster III, and cluster V, it was harder to match them up with the remaining dialects because of slight inconsistencies between the heat maps and the linguistic atlas.

Again, because of the ambiguity between latent class analysis results and the heat maps, I incorporated data from the atlas to interpret the results. For the Northern dialect, the atlas marks the Northern Cities Chain shift that affects lax vowels as a central distinguishing factor for this dialect. The survey variables "been" and "bag" can be markers for this feature. The highest probabilities of 'been' pronounced as /ɪ/ and 'bag' pronounced as /eɪ/ occur in cluster I. This is evidence that cluster I can be classified as the North dialect even though the Midland dialect had the most points for cluster I. Still, based on results from the linguistic atlas marking the Northern Cities Chain Shift as a major distinguishing characteristic of the North dialect, I labeled cluster I as the North dialect, consistent with both the data from my findings and the linguistic atlas.

Similarly, the Midland dialect, the West dialect, and the North dialect had high points for cluster III and cluster V. However, because I already classified the North as cluster 1, I discarded the North dialect as a possible label for cluster III or cluster V, leaving only the Midland dialect or the West dialect as a possible label. Again, I looked to the linguistic atlas and aligned its findings with my data. The Atlas lists the low back merger ("cot/caught") as a distinguishing feature of the West dialect. In the Midland dialect, this merger is transitional. From this

information, I posit that cluster III represents the West dialect and cluster V represents the Midland dialect because cluster III has a slightly lower probability than cluster V for "cot" and "caught" to be pronounced differently.

With these considerations in mind, I conclude that the clusters from the latent class analysis represent the following dialects:

Cluster I	The North
Cluster II	The South
Cluster III	The West
Cluster IV	New England
Cluster V	The Midland
Cluster VI	New York City and the Mid-Atlantic States

However, some variables in the clusters do not "fit" with their designated dialect. For example, "pajamas" has two pronunciations: "paj/ɑ/mas" and "paj/æ/mas". According to Katz's heat maps of the distribution, "paj/æ/mas" is more prevalent in the West dialect, the North dialect, and the Midland dialect, while the pronunciation "paj/ɑ/mas" is more common in the South dialect, New York City and the Mid-Atlantic States dialect, and areas of the New England dialect. However, the clusters are arranged so that four clusters have a very high probability of "paj/ɑ/mas" with a corresponding low probability of "paj/æ/mas" and two clusters have a high probability of "paj/æ/mas" with a low probability of "paj/ɑ/mas" (Tables 2, 3, 4, 5, 6, 7). Because of the discrepancy between the cluster output and the heat maps, it is unavoidable that there is a dialect where the "pajamas" variable does not fit perfectly. In the way that I have labeled the clusters, the incongruity with "pajamas" occurs in the West dialect.

LCA Results for Dialects

The results from the latent class analysis can be seen in Table 2, 3, 4, 5, 6, and 7.

THE NORTH

Table 2: Latent class analysis results for the North dialect.

CLUSTER 1: The North			
pajamas: /æ/ as in "jam"	0.984	caramel: 3 syllables “car-ra-mel”	0.313
coupon: /ju:/ as in “cute”	0.982	Craig: in between /ɛ/ and /e:/	0.288
Monday, Friday: /e:/ as in "say"	0.934	flourish: /ɔ/ as in "sore"	0.269
Florida: /ɔ/ as in "sore"	0.900	handkerchief: /i:/ as in "see"	0.239
caramel: 2 syllables “car-ml”	0.876	creek: /ɪ/ as in "sit"	0.228
lawyer: /ɔj/ as in "boy"	0.841	crayon: /æ/ as in “man”	0.219
route: /raʊt/ rhymes with "out"	0.794	pecan: /pi:kɑn/ "PEE-kahn"	0.211
cauliflower: /ɪ/ as in “sit”	0.763	syrup: /i:/	0.208
bowie knife: /o:/ as in “bo”	0.748	lawyer: /ɒ/ as in "saw"	0.200
handkerchief: /ɪ/ as in "sit"	0.743	Craig: /ɛ/ as in “set”	0.194
flourish: /ɚ/ as in "bird"	0.659	cauliflower: /i:/ as in “see”	0.172
syrup: /ɚ/	0.646	bag: /e:/	0.159
miracle: /i:/ as in "near"	0.645	bowie knife: /u:/ as in “boo”	0.153
route: /ru:t/ rhymes with "hoot"	0.631	syrup: /ɪ/	0.126
poem: 2 syllables	0.590	aunt: /ɑ/ as in “ah”	0.121
mayonnaise: /æ/ as in "man" (2 syl)	0.544	pecan: /pí:kæn/ "PEE-can"	0.098
been: /ɪ/ as in “sit”	0.518	Monday, Friday: /i:/ as in "see"	0.095
Craig: /e:/ as in “say”	0.509	really: /iə/ "ree-l-y"	0.071
mayonnaise: /ejɑ/ (3 syl.)	0.492	aunt: /ɒ/ as in “caught”	0.070
really: /i:/ as in "see"	0.483	pecan: /pi:ḱæn/ "pee-CAN"	0.066
cot ≠caught (/ɑ/ and /ɔ/)	0.436	Florida: /o:/ as in "flow"	0.050
realtor: 3 syllables (with /ə/)	0.433	flourish: /ʌ/ as in "sun"	0.035
been: /ɛ/ as in “set”	0.422	been: /i:/ as in “see”	0.026
poem: 1 syllable	0.405	Florida: /ɒ/ as in "saw"	0.013
crayon: /ejɒ/ (2 syl, "cray-awn")	0.372	Florida: /ɑ/ as in "ah"	0.012
realtor: 2 syllables	0.369	coupon: /u:/ as in “coop”	0.000
crayon: /ejɑ/ (2 syl, "cray-ahn")	0.357	pajamas: /ɑ/ as in "father"	0.000

*The left column represents the linguistic variables and pronunciations and the right column represents the probability each variable will occur in the cluster.

Cluster I most closely resembles the North dialect with 19 matches (Table 1). For example, the results from the latent class analysis show the North dialect with the highest probability for the vowel in ‘bag’ pronounced with an /e/. This is strong evidence of the presence of the North Cities Shift. The latent class analysis also paralleled the linguistic atlas with the

pronunciation of ‘Florida’ with the low back vowel /ɔ/ showing the absence of the low back vowel merger with this particular variable.

Although cluster I most accurately resembles the North dialect, there are some problematic variables. For example, the heat maps show the variable “mayonnaise” to have a higher pronunciation of “m/ejɑ/nnaise” than “m/æ/nnaise”. However, the opposite is true in cluster I. Also, perhaps a more significant variable is the low back vowel merger. One major discrepancy between the latent class analysis and the linguistic atlas is the variable with the low back merger found within the variable ‘cot/caught’. The linguistic atlas states that the two low back vowels /o/ and /ɔ/ are distinct in the North, especially in the Inland North region. However, the results from the latent class analysis show the North as having not only a low percentage of the low back merger, but the lowest (0.436) out of all the six dialects of American English.

THE SOUTH

Table 3: Latent class analysis results for the South dialect.

CLUSTER 2: The South			
handkerchief: /ɪ/ as in "sit"	0.922	really: /i:/ as in "see"	0.295
lawyer: /ɒ/ as in "saw"	0.903	realtor: 3 syllables (with /ə/)	0.260
pajamas: /ɑ/ as in "father"	0.902	crayon: /ejɒ/ (2 syl, "cray-awn")	0.240
poem: 2 syllables	0.864	Craig: /e:/ as in “say”	0.236
route: /raʊt/ rhymes with "out"	0.794	Florida: /ɑ/ as in "ah"	0.207
cauliflower: /ɪ/ as in “sit”	0.784	been: /ɛ/ as in “set”	0.176
been: /ɪ/ as in “sit”	0.767	syrup: /ɪ/	0.157
cot ≠caught (/ɑ/ and /ɔ/)	0.755	lawyer: /ɔj/ as in "boy"	0.141
flourish: /ɚ/ as in "bird"	0.749	flourish: /ɔ/ as in "sore"	0.134
Monday, Friday: /e:/ as in "say"	0.720	cauliflower: /i:/ as in “see”	0.132
caramel: 3 syllables “car-ra-mel”	0.711	poem: 1 syllable	0.128
crayon: /ejɑ/ (2 syl, "cray-ahn")	0.708	syrup: /i:/	0.115
syrup: /ɚ/	0.703	really: /iə/ "ree-l-y"	0.114
route: /ru:t/ rhymes with "hoot"	0.676	Florida: /ɒ/ as in "saw"	0.097
Florida: /ɔ/ as in "sore"	0.638	aunt: /ɑ/ as in “ah”	0.085
mayonnaise: /ejɑ/ (3 syl.)	0.562	flourish: /ʌ/ as in "sun"	0.081
coupon: /ju:/ as in “cute”	0.555	pajamas: /æ/ as in "jam"	0.075
realtor: 2 syllables	0.514	creek: /ɪ/ as in "sit"	0.070
bowie knife: /u:/ as in “boo”	0.507	pecan: /pí:kæn/ "PEE-can"	0.064
mayonnaise: /æ/ as in "man" (2 syl)	0.505	pecan: /pi:kɑn/ "PEE-kahn"	0.061
miracle: /i:/ as in "near"	0.487	handkerchief: /i:/ as in "see"	0.057
Craig: /ɛ/ as in “set”	0.449	pecan: /pi:kǽn/ "pee-CAN"	0.046
bowie knife: /o:/ as in “bo”	0.439	aunt: /ɒ/ as in “caught”	0.045
coupon: /u:/ as in “coop”	0.424	Florida: /o:/ as in "flow"	0.028
caramel: 2 syllables “car-ml”	0.416	bag: /e:/	0.025
Monday, Friday: /i:/ as in "see"	0.382	crayon: /æ/ as in “man”	0.019
Craig: in between /ɛ/ and /e:/	0.302	been: /i:/ as in “see”	0.018

*The left column represents the linguistic variables and pronunciations and the right column represents the probability each variable will occur in the cluster.

The results from the latent class analysis found the South to most closely match cluster II with 22 matches (Table 1). The results show the South to have a strong pronunciation of ‘lawyer’ using the vowel /a/ instead of the diphthong. The analysis also found the South to have the second strongest probability for the low back vowel distinction in ‘cot/caught’ with a probability

of 0.755. Also the South as the highest probability (0.864) of ‘poem’ being pronounced with a diphthong, or two syllables. This is similar to the findings in the linguistic atlas where long vowels are broken up with a glide.

However, even though cluster II corresponds with the South dialect, some variables do not parallel with the pattern. For instance, “mayonnaise” should have a higher pronunciation of “m/æ/nnaise” than “m/ejɑ/nnaise”; however, the opposite is true. The latent class analysis reports a larger probability of the pronunciation with the diphthong /ejɑ/ (0.562) instead of /æ/ (0.505). This is the only variable for this cluster that does not fit with the label of the South, even though this is only a slight difference.

THE WEST

Table 4: Latent class analysis of the West dialect.

CLUSTER 3: The West

pajamas: /ɑ/ as in "father"	0.970	flourish: /ɔ/ as in "sore"	0.325
Florida: /ɔ/ as in "sore"	0.967	been: /ɛ/ as in “set”	0.317
Monday, Friday: /e:/ as in "say"	0.967	realtor: 3 syllables (with /ə/)	0.313
handkerchief: /ɪ/ as in "sit"	0.878	Craig: /ɛ/ as in “set”	0.272
lawyer: /ɔj/ as in "boy"	0.870	crayon: /ejɑ/ (2 syl, "cray-ahn")	0.264
cauliflower: /ɪ/ as in “sit”	0.835	lawyer: /ɒ/ as in "saw"	0.203
route: /ru:t/ rhymes with "hoot"	0.802	aunt: /ɑ/ as in “ah”	0.200
bowie knife: /o:/ as in “bo”	0.775	syrup: /i:/	0.196
route: /raʊt/ rhymes with "out"	0.765	syrup: /ɪ/	0.144
caramel: 2 syllables “car-ml”	0.738	bowie knife: /u:/ as in “boo”	0.142
mayonnaise: /æ/ as in "man" (2 syl)	0.663	aunt: /ɒ/ as in “caught”	0.139
poem: 2 syllables	0.659	creek: /ɪ/ as in "sit"	0.128
been: /ɪ/ as in “sit”	0.640	cauliflower: /i:/ as in “see”	0.112
syrup: /ɚ/	0.634	crayon: /æ/ as in “man”	0.112
caramel: 3 syllables “car-ra-mel”	0.614	pecan: /pi:kɑn/ "PEE-kahn"	0.112
miracle: /i:/ as in "near"	0.603	handkerchief: /i:/ as in "see"	0.105
flourish: /ɚ/ as in "bird"	0.602	really: /iə/ "ree-l-y"	0.092
crayon: /ejɒ/ (2 syl, "cray-awn")	0.589	pecan: /pí:kæn/ "PEE-can"	0.062
coupon: /u:/ as in “coop”	0.574	Monday, Friday: /i:/ as in "see"	0.060
cot ≠caught (/ɑ/ and /ɔ/)	0.486	pecan: /pi:kǽn/ "pee-CAN"	0.042
really: /i:/ as in "see"	0.433	bag: /e:/	0.040
coupon: /ju:/ as in “cute”	0.408	flourish: /ʌ/ as in "sun"	0.030
mayonnaise: /ejɑ/ (3 syl.)	0.402	Florida: /o:/ as in "flow"	0.017
Craig: in between /ɛ/ and /e:/	0.379	been: /i:/ as in “see”	0.015
realtor: 2 syllables	0.364	Florida: /ɑ/ as in "ah"	0.000
Craig: /e:/ as in “say”	0.340	Florida: /ɒ/ as in "saw"	0.000
poem: 1 syllable	0.336	pajamas: /æ/ as in "jam"	0.000

*The left column represents the linguistic variables and pronunciations and the right column represents the probability each variable will occur in the cluster.

Cluster 3 most closely resembles the West dialect with 21 matches (Table 1). The findings from the latent class analysis show the distinction between the low back vowels in ‘cot/caught’ to be the second lowest (0.486) to the North (0.436). However, the West does show a high probability of the second vowel in ‘pajamas’ to be pronounced with /ɑ/ and not /æ/ as the heat maps would have predicted.

While cluster III most closely resembles the West, some variables that do not fit exactly with this label include “route” and “pajamas” as mentioned previously. According to the heat maps, “route” should have a higher pronunciation of “r/aʊ/t”. Yet, cluster III has a higher probability of the “r/u:/t” pronunciation.

Table 5: Latent class analysis of New England dialect.

CLUSTER 4: New England

Monday, Friday: /e:/ as in "say"	0.833	pajamas: /æ/ as in "jam"	0.332
route: /ru:t/ rhymes with "hoot"	0.755	lawyer: /ɒ/ as in "saw"	0.319
caramel: 3 syllables “car-ra-mel”	0.729	been: /ɛ/ as in “set”	0.318
mayonnaise: /ejɑ/ (3 syl.)	0.712	realtor: 2 syllables	0.317
lawyer: /ɔj/ as in "boy"	0.664	pecan: /pí:kæn/ "PEE-can"	0.308
bowie knife: /o:/ as in “bo”	0.662	coupon: /ju:/ as in “cute”	0.300
poem: 2 syllables	0.661	flourish: /ɔ/ as in "sore"	0.286
coupon: /u:/ as in “coop”	0.642	poem: 1 syllable	0.271
handkerchief: /i:/ as in "see"	0.568	mayonnaise: /æ/ as in "man" (2 syl)	0.255
Craig: /e:/ as in “say”	0.566	syrup: /ɪ/	0.254
really: /i:/ as in "see"	0.563	bowie knife: /u:/ as in “boo”	0.219
pajamas: /ɑ/ as in "father"	0.558	flourish: /ʌ/ as in "sun"	0.207
cot ≠caught (/ɑ/ and /ɔ/)	0.555	Craig: in between /ɛ/ and /e:/	0.204
miracle: /i:/ as in "near"	0.530	syrup: /ɚ/	0.200
cauliflower: /i:/ as in “see”	0.515	really: /iə/ "ree-l-y"	0.176
Florida: /ɔ/ as in "sore"	0.502	Craig: /ɛ/ as in “set”	0.169
route: /raʊt/ rhymes with "out"	0.469	Florida: /o:/ as in "flow"	0.166
been: /ɪ/ as in “sit”	0.454	been: /i:/ as in “see”	0.140
syrup: /i:/	0.450	Florida: /ɑ/ as in "ah"	0.134
crayon: /ejɑ/ (2 syl, "cray-ahn")	0.438	Florida: /ɒ/ as in "saw"	0.131
crayon: /ejɒ/ (2 syl, "cray-awn")	0.431	aunt: /ɒ/ as in “caught”	0.129
flourish: /ɚ/ as in "bird"	0.409	pecan: /pi:kɑn/ "PEE-kahn"	0.107
cauliflower: /ɪ/ as in “sit”	0.406	pecan: /pi:ḱæn/ "pee-CAN"	0.103
caramel: 2 syllables “car-ml”	0.400	bag: /e:/	0.096
aunt: /ɑ/ as in “ah”	0.349	Monday, Friday: /i:/ as in "see"	0.066
realtor: 3 syllables (with /ə/)	0.349	creek: /ɪ/ as in "sit"	0.062
handkerchief: /ɪ/ as in "sit"	0.342	crayon: /æ/ as in “man”	0.061

*The left column represents the linguistic variables and pronunciations and the right column represents the probability each variable will occur in the cluster.

Cluster 4 closely resembles the New England dialect with 20 matches (Table 1). The latent class analysis shows New England’s pronunciation of ‘aunt’ with /ɑ/ with the highest probability (0.349) of the dialects. However, the pronunciation of the second vowel in ‘pajamas’ with the same vowel /ɑ/ does not have a very high or relatively high probability compared with the other dialects that the linguistic atlas would predict. The probability is 0.558 and is the fourth

highest dialect probability. The distinction of the vowels in ‘cot/caught’ is also moderate at 0.555 probability of no merger in the dialect region. This may be a result of the New England dialect itself being split north to south on the low back merger.

Table 6: Latent class analysis of the Midland dialect

CLUSTER 5: The Midland			
pajamas: /æ/ as in "jam"	0.986	crayon: /ejɒ/ (2 syl, "cray-awn")	0.335
coupon: /u:/ as in “coop”	0.983	Craig: in between /ɛ/ and /e:/	0.286
Monday, Friday: /e:/ as in "say"	0.962	syrup: /i:/	0.244
Florida: /ɔ/ as in "sore"	0.896	handkerchief: /i:/ as in "see"	0.241
lawyer: /ɔj/ as in "boy"	0.879	flourish: /ɔ/ as in "sore"	0.240
caramel: 2 syllables “car-ml”	0.805	Craig: /ɛ/ as in “set”	0.237
Bowie knife: /o:/ as in “bo”	0.776	crayon: /æ/ as in “man”	0.235
route: /raʊt/ rhymes with "out"	0.764	cauliflower: /i:/ as in “see”	0.203
handkerchief: /ɪ/ as in "sit"	0.739	pecan: /pi:kɑn/ "PEE-kahn"	0.188
cauliflower: /ɪ/ as in “sit”	0.736	creek: /ɪ/ as in "sit"	0.180
route: /ru:t/ rhymes with "hoot"	0.704	lawyer: /ɒ/ as in "saw"	0.163
flourish: /ɚ/ as in "bird"	0.680	aunt: /ɑ/ as in “ah”	0.140
miracle: /i:/ as in "near"	0.644	syrup: /ɪ/	0.138
poem: 2 syllables	0.598	bowie knife: /u:/ as in “boo”	0.129
syrup: /ɚ/	0.597	pecan: /pí:kæn/ "PEE-can"	0.115
mayonnaise: /æ/ as in "man" (2 syl)	0.570	aunt: /ɒ/ as in “caught”	0.089
been: /ɪ/ as in “sit”	0.537	pecan: /pi:kán/ "pee-CAN"	0.071
really: /i:/ as in "see"	0.509	Monday, Friday: /i:/ as in "see"	0.069
cot ≠caught (/ɑ/ and /ɔ/)	0.507	really: /iə/ "ree-l-y"	0.062
mayonnaise: /ejɑ/ (3 syl.)	0.485	Florida: /o:/ as in "flow"	0.060
Craig: /e:/ as in “say”	0.467	flourish: /ʌ/ as in "sun"	0.034
realtor: 3 syllables (with /ə/)	0.411	been: /i:/ as in “see”	0.028
been: /ɛ/ as in “set”	0.406	Florida: /ɑ/ as in "ah"	0.012
caramel: 3 syllables “car-ra-mel”	0.401	Florida: /ɒ/ as in "saw"	0.010
crayon: /ejɑ/ (2 syl, "cray-ahn")	0.401	bag: /e:/	0.006
poem: 1 syllable	0.397	coupon: /ju:/ as in “cute”	0.000
realtor: 2 syllables	0.381	pajamas: /ɑ/ as in "father"	0.000

*The left column represents the linguistic variables and pronunciations and the right column represents the probability each variable will occur in the cluster.

The results in cluster V most closely resemble the Midland dialect with 20 matches with the heat maps (Table 1). In the latent class analysis, the Midland has an average probability for the low back merger. This can be expected based on the linguistic atlas since the Midland is a transitional zone for the low back merger.

With cluster V labeled as the Midland dialect, the variable “lawyer” does not correspond correctly with the dialect. The heat maps show the Midland dialect as having both “l/ɔj/er” and “l/ɑ/yer” as possible pronunciations for “lawyer”. However, cluster V has a high probability of the “l/ɔj/er” pronunciation (0.879) and a low probability for the “l/ɑ/yer” pronunciation (0.163).

Table 7: Latent class analysis of New York City and the Mid-Atlantic States dialect.

CLUSTER 6: New York City & the Mid-Atlantic States

handkerchief: /ɪ/ as in "sit"	0.995	Florida: /ɑ/ as in "ah"	0.310
Monday, Friday: /e:/ as in "say"	0.945	mayonnaise: /æ/ as in "man" (2 syl)	0.306
pajamas: /ɑ/ as in "father"	0.927	poem: 1 syllable	0.303
been: /ɪ/ as in “sit”	0.917	aunt: /ɑ/ as in “ah”	0.289
route: /ru:t/ rhymes with "hoot"	0.906	pecan: /pí:kæn/ "PEE-can"	0.270
lawyer: /ɔj/ as in "boy"	0.880	bowie knife: /u:/ as in “boo”	0.227
coupon: /u:/ as in “coop”	0.876	flourish: /ʌ/ as in "sun"	0.191
cot ≠caught (/ɑ/ and /ɔ/)	0.828	realtor: 3 syllables (with /ə/)	0.188
caramel: 3 syllables “car-ra-mel”	0.765	Florida: /ɒ/ as in "saw"	0.184
crayon: /ejɑ/ (2 syl, "cray-ahn")	0.752	crayon: /ejɒ/ (2 syl, "cray-awn")	0.176
mayonnaise: /ejɑ/ (3 syl.)	0.718	syrup: /ɚ/	0.175
poem: 2 syllables	0.692	lawyer: /ɒ/ as in "saw"	0.160
really: /i:/ as in "see"	0.685	pecan: /pi:ḱən/ "pee-CAN"	0.129
bowie knife: /o:/ as in “bo”	0.666	flourish: /ɔ/ as in "sore"	0.128
flourish: /ɚ/ as in "bird"	0.645	aunt: /ɒ/ as in “caught”	0.118
realtor: 2 syllables	0.550	coupon: /ju:/ as in “cute”	0.112
cauliflower: /i:/ as in “see”	0.529	really: /iə/ "ree-l-y"	0.110
miracle: /i:/ as in "near"	0.496	creek: /ɪ/ as in "sit"	0.070
route: /raʊt/ rhymes with "out"	0.456	Monday, Friday: /i:/ as in "see"	0.070
Florida: /ɔ/ as in "sore"	0.450	pecan: /pi:kɑn/ "PEE-kahn"	0.061
cauliflower: /ɪ/ as in “sit”	0.443	been: /ɛ/ as in “set”	0.057
syrup: /i:/	0.402	pajamas: /æ/ as in "jam"	0.054
syrup: /ɪ/	0.396	crayon: /æ/ as in “man”	0.036
Craig: /e:/ as in “say”	0.357	bag: /e:/	0.031
caramel: 2 syllables “car-ml”	0.324	Florida: /o:/ as in "flow"	0.029
Craig: in between /ɛ/ and /e:/	0.319	been: /i:/ as in “see”	0.009
Craig: /ɛ/ as in “set”	0.314	handkerchief: /i:/ as in "see"	0.000

*The left column represents the linguistic variables and pronunciations and the right column represents the probability each variable will occur in the cluster.

Cluster VI most closely resembles the New York City and the Mid-Atlantic States dialect region with 24 matches to the heat maps (Table 1). The results from the latent class analysis show the highest probability of the distinction between the low back vowels in ‘cot/caught’ with

a probability of 0.828. This parallels the findings from both the linguistic atlas and the heat maps. Also, the probability of the pronunciation of 'pajamas' with /ɑ/ (0.927) is higher than the pronunciation with /æ/ (0.054). Similarly, the pronunciation of 'syrup' with /i/ or /ɪ/ (0.402 and 0.396 respectively) were higher than the /ə/ pronunciation (0.175). Unfortunately, due to the nature of the questions, there were no questions eliciting for r-pronunciation or the short-a split.

After labeling the clusters, I created a chart showing each variable ranked by its probability of occurring in each cluster from highest to lowest. The cluster labels are also included in the chart. This chart can be seen in Table 8.

TABLE 8: Dialects for each variable ranked by its probability from the latent class analysis.

aunt: /ɑ/ as in "ah"			Bowie knife: /u:/ as in "boo"			coupon: /u:/ as in "coop"		
IV	0.349	N. England	II	0.507	South	V	0.983	Midland
VI	0.289	Mid-Atlantic	VI	0.227	Mid-Atlantic	VI	0.876	Mid-Atlantic
III	0.200	West	IV	0.219	N. England	IV	0.642	N. England
V	0.140	Midland	I	0.153	North	III	0.574	West
I	0.121	North	III	0.142	West	II	0.424	South
II	0.085	South	V	0.129	Midland	I	0.000	North
aunt: /ɒ/ as in "caught"			caramel: 2 syl. "car-ml"			coupon: /ju:/ as in "cute"		
III	0.139	West	I	0.876	North	I	0.982	North
IV	0.129	N. England	V	0.805	Midland	VI	0.888	Mid-Atlantic
VI	0.118	Mid-Atlantic	III	0.738	West	II	0.555	South
V	0.089	Midland	II	0.416	South	III	0.408	West
I	0.070	North	IV	0.400	N. England	IV	0.300	N. England
II	0.045	South	VI	0.324	Mid-Atlantic	V	0.000	Midland
Been: /ɪ/ as in "sit"			caramel: 3 syl. "car-ra-mel"			Craig: /ɛ/ as in "set"		
VI	0.917	Mid-Atlantic	VI	0.765	Mid-Atlantic	II	0.449	South
II	0.767	South	IV	0.729	N. England	VI	0.314	Mid-Atlantic
III	0.640	West	II	0.711	South	III	0.272	West
V	0.537	Midland	III	0.614	West	V	0.237	Midland
I	0.518	North	V	0.401	Midland	I	0.194	North
IV	0.454	N. England	I	0.313	North	IV	.0169	N. England
been: /i:/ as in "see"			cauliflower: /i:/ as in "see"			Craig: /e:/ as in "say"		
IV	0.140	N. England	VI	0.529	Mid-Atlantic	IV	0.566	N. England
V	0.028	Midland	IV	0.515	N. England	I	0.509	North
I	0.026	North	V	0.203	Midland	V	0.467	Midland
II	0.018	South	I	0.172	North	VI	0.357	Mid-Atlantic
III	0.015	West	II	0.132	South	III	0.340	West
VI	0.009	Mid-Atlantic	III	0.112	West	II	0.236	South
Bowie knife: /o:/ as in "bo"			cauliflower: /ɪ/ as in "sit"			Craig: between /ɛ/ and /e:/		
V	0.776	Midland	III	0.835	West	III	0.379	West
III	0.775	West	II	0.784	South	VI	0.319	Mid-Atlantic
I	0.748	North	I	0.763	North	II	0.302	South
VI	0.666	Mid-Atlantic	V	0.736	Midland	I	0.288	North
IV	0.662	N. England	VI	0.443	Mid-Atlantic	V	0.286	Midland
II	0.439	South	IV	0.406	N. England	IV	0.204	N. England

crayon: /æ/ as in "man"

V	0.235	Midland
I	0.219	North
III	0.112	West
IV	0.061	N. England
VI	0.036	Mid-Atlantic
II	0.019	South

crayon: /ejɑ/ 2 syl, "cray-ahn"

VI	0.752	Mid-Atlantic
II	0.708	South
IV	0.438	N. England
V	0.401	Midland
I	0.357	North
III	0.264	West

crayon: /ejɒ/ 2 syl, "cray-awn"

III	0.589	West
IV	0.431	N. England
I	0.372	North
V	0.335	Midland
II	0.240	South
VI	0.176	Mid-Atlantic

creek: /ɪ/ as in "sit"

I	0.228	North
V	0.180	Midland
III	0.128	West
II	0.070	South
VI	0.070	Mid-Atlantic
IV	0.062	N. England

Florida: /oː/ as in "flow"

IV	0.166	N. England
V	0.060	Midland
I	0.050	North
VI	0.029	Mid-Atlantic
II	0.028	South
III	0.017	West

Florida: /ɑ/ as in "ah"

VI	0.310	Mid-Atlantic
II	0.207	South
IV	0.131	N. England
I	0.012	North
V	0.060	Midland
III	0.000	West

Florida: /ɒ/ as in "saw"

VI	0.184	Mid-Atlantic
IV	0.131	N. England
II	0.097	South
I	0.013	North
V	0.010	Midland
III	0.000	West

Florida: /ɔ/ as in "sore"

III	0.967	West
I	0.900	North
V	0.896	Midland
II	0.638	South
IV	0.502	N. England
VI	0.450	Mid-Atlantic

flourish: /ɚ/ as in "bird"

II	0.749	South
V	0.680	Midland
I	0.659	North
VI	0.645	Mid-Atlantic
III	0.602	West
IV	0.502	N. England

flourish: /ɔ/ as in "sore"

III	0.325	West
IV	0.286	N. England
I	0.269	North
V	0.240	Midland
II	0.134	South
VI	0.128	Mid-Atlantic

flourish: /ʌ/ as in "sun"

IV	0.207	N. England
VI	0.191	Mid-Atlantic
II	0.081	South
I	0.035	North
V	0.034	Midland
III	0.030	West

handkerchief: /iː/ as in "see"

IV	0.568	N. England
V	0.241	Midland
I	0.239	North
III	0.105	West
II	0.057	South
VI	0.000	Mid-Atlantic

handkerchief: /ɪ/ as in "sit"

VI	0.995	Mid-Atlantic
II	0.922	South
III	0.878	West
I	0.743	North
V	0.739	Midland
IV	0.342	N. England

lawyer: /ɔj/ as in "boy"

VI	0.880	Mid-Atlantic
V	0.879	Midland
III	0.870	West
I	0.841	North
IV	0.664	N. England
II	0.141	South

lawyer: /ɒ/ as in "saw"

II	0.903	South
IV	0.319	N. England
III	0.203	West
I	0.200	North
V	0.163	Midland
VI	0.160	Mid-Atlantic

mayonnaise: /æ/ 2 syl.

III	0.663	West
V	0.570	Midland
I	0.544	North
II	0.505	South
VI	0.306	Mid-Atlantic
IV	0.255	N. England

mayonnaise: /ejɑ/ 3 syl.

VI	0.718	Mid-Atlantic
IV	0.712	N. England
II	0.562	South
I	0.492	North
V	0.485	Midland
III	0.402	West

miracle: /i:/ as in "near"

I	0.645	North
V	0.644	Midland
III	0.603	West
IV	0.530	N. England
VI	0.496	Mid-Atlantic
II	0.487	South

Monday, Friday: /e:/ as in "say"

III	0.967	West
V	0.962	Midland
VI	0.945	Mid-Atlantic
I	0.934	North
IV	0.833	N. England
II	0.720	South

Monday, Friday: /i:/ as in "see"

II	0.382	South
I	0.095	North
VI	0.070	Mid-Atlantic
V	0.069	Midland
IV	0.066	N. England
III	0.060	West

pajamas: /æ/ as in "jam"

V	0.986	Midland
I	0.984	North
IV	0.332	N. England
II	0.075	South
VI	0.054	Mid-Atlantic
III	0.000	West

pajamas: /ɑ/ as in "father"

III	0.970	West
VI	0.927	Mid-Atlantic
II	0.902	South
IV	0.558	N. England
I	0.000	North
V	0.000	Midland

pecan: /pí:kæn/ "PEE-can"

IV	0.308	N. England
VI	0.270	Mid-Atlantic
V	0.115	Midland
I	0.098	North
II	0.064	South
III	0.062	West

pecan: /pi:kǽn/ "pee-CAN"

6	0.129	Mid-Atlantic
4	0.103	N. England
5	0.071	Midland
1	0.066	North
2	0.046	South
3	0.042	West

pecan: /pi:kɑ n/ "pee-KAHN"

I	0.211	North
V	0.188	Midland
III	0.112	West
IV	0.107	N. England
II	0.061	South
VI	0.061	Mid-Atlantic

poem: 1 syllable

I	0.405	North
V	0.397	Midland
III	0.336	West
VI	0.303	Mid-Atlantic
IV	0.271	N. England
II	0.128	South

poem: 2 syllables

II	0.864	South
VI	0.692	Mid-Atlantic
IV	0.661	N. England
III	0.659	West
V	0.598	Midland
I	0.590	North

really: /i:/ as in "see"

VI	0.685	Mid-Atlantic
IV	0.563	N. England
V	0.509	Midland
I	0.483	North
III	0.433	West
II	0.295	South

really: /iə/ "ree-l-y"

4	0.176	N. England
2	0.114	South
6	0.110	Mid-Atlantic
3	0.092	West
1	0.071	North
5	0.062	Midland

realtor: 2 syllables

VI	0.550	Mid-Atlantic
II	0.514	South
V	0.381	Midland
I	0.369	North
III	0.364	West
IV	0.317	N. England

realtor: 3 syllables (with /ə/)

I	0.433	North
V	0.411	Midland
IV	0.349	N. England
III	0.313	West
II	0.260	South
VI	0.188	Mid-Atlantic

route: /ru:t/ rhymes with "hoot"

VI	0.906	Mid-Atlantic
III	0.802	West
IV	0.755	N. England
V	0.704	Midland
II	0.676	South
I	0.631	North

route: /raʊt/ rhymes with "out"

I	0.794	North
II	0.794	South
III	0.765	West
V	0.764	Midland
IV	0.469	N. England
VI	0.456	Mid-Atlantic

syrup: /i:/

IV	0.450	N. England
VI	0.402	Mid-Atlantic
V	0.244	Midland
I	0.208	North
III	0.196	West
II	0.115	South

syrup: /ɪ/

VI	0.396	Mid-Atlantic
IV	0.254	N. England
II	0.157	South
III	0.144	West
V	0.138	Midland
I	0.126	North

syrup: /ɚ/

II	0.703	South
I	0.646	North
III	0.634	West
V	0.597	Midland
IV	0.200	N. England
VI	0.175	Mid-Atlantic

cot≠ caught (/ɑ/ and /ɔ/)

VI	0.828	Mid-Atlantic
II	0.755	South
IV	0.555	N. England
V	0.507	Midland
III	0.486	West
I	0.436	North

bag: /e:/

I	0.159	North
V	0.148	Midland
IV	0.096	N. England
III	0.040	West
VI	0.031	Mid-Atlantic
II	0.025	South

Correlation Analysis

TABLE 9: Correlation analysis showing the similarities of each dialect to the other dialects by r values

	I. North	II. South	III. West	IV. New England	V. Midland	VI. Mid-Atlantic States and NYC
I. North	1.000	0.480**	0.684**	0.492**	0.786**	0.445**
II. South	0.480**	1.000	0.731**	0.503**	0.463**	0.684**
III. West	0.684**	0.731**	1.000	0.677**	0.727**	0.712**
IV. New England	0.492**	0.503**	0.677**	1.000	0.636**	0.741**
V. Midland	0.786**	0.463**	0.727**	0.636**	1.000	0.493**
VI. Mid-Atlantic States and NYC	0.445**	0.684**	0.712**	0.741**	0.493**	1.000

The correlation analysis shows the Pearson Correlation Coefficient. ** represents significance at the 0.01 level.

The results from the correlation analysis based on r values (Table 9) show how closely each dialect is similar to the others based on the probabilities from the latent class analysis. As shown in Table 9, for the North dialect, the dialect that was most similar is the Midland, while the New York City and the Mid-Atlantic States dialect is the most different. The South dialect was most similar to the West dialect and the most different from the Midland dialect. The West dialect was most similar to the South and most different from the New England dialect. The New England dialect was most similar to the New York City and the Mid-Atlantic States dialect while most different from the North dialect. The Midland was most similar to the North and most different from the South dialect. And finally, the New York City and the Mid-Atlantic States dialect was most similar to the New England dialect and most different than the North. All correlations were found to be significant at the 0.01 level except for the correlations where the dialects are compared to themselves.

The dialects that had the largest range between how they correlated with other dialects was the North dialect and the New York City and the Mid-Atlantic States dialect both with a range of 0.555. The West, on the other hand, was the dialect that had the smallest range in its correlations with other dialects at 0.323.

DISCUSSION AND CONCLUSION

In this analysis of American English dialect regions, I used a latent class analysis to generate six dialects whose linguistic features naturally occurred together. In each dialect, I found the probability of each variable from the Harvard Dialect Survey occurring in the dialect. This allowed me to answer my first research question of discovering what phonetic variables from the Harvard Dialect Survey are most closely associated with each dialect.

For the North dialect, the variables with a probability greater than 0.850 were 'pajamas: /æ/' (0.984), 'coupon: /ju:/' (0.982), 'Monday, Friday: /e:/' (0.934), 'Florida: /ɔ/' (0.900), and 'caramel: 2 syllables' (0.876). The variable eliciting pronunciation of the vowel in 'bag' with /e:/ had the highest probability of occurring in the North at 0.159. Even though this value is low, it was still the highest of all the dialects. And the probability of the distinction between the two vowels in the low back merger in 'cot/caught' had the lowest probability of occurring in the North at 0.436.

In the South dialect, the variables that had a probability greater of 0.850 were 'handkerchief: /ɪ/' (0.922), 'lawyer: /ɒ/' (0.903), 'pajamas: /ɑ/' (0.902), and 'poem' as 2 syllables. The 'cot/caught' distinction between /ɑ/ and /ɔ/ had a probability of 0.755 of occurring.

The phonetic variables with the highest probability of the West dialect included 'pajamas: /ɑ/' (0.970), 'Florida: /ɔ/' (0.967), 'Monday, Friday: /e:/' (0.967), 'handkerchief: /ɪ/' (0.878), and 'lawyer: /ɔj/' (0.870). The low back vowel distinction in 'cot/caught' was low with a probability of 0.486.

For the New England dialect, no variables occurred at a probability of greater than 0.850. However, the top five variables for the dialect include 'Monday, Friday: /e:/' (0.833), 'route:

/ru:t/' (0.755), 'caramel: 3 syllables' (0.729), 'mayonnaise: /ejɑ/' (0.712) and 'lawyer: /ɔj/' (0.664). The 'cot/caught' distinction occurred at a probability of 0.555.

The Midland dialect top phonetic variables were 'pajamas: /æ/' (0.986), 'coupon: /u:/' (0.983), 'Monday, Friday: /e:/' (0.962), 'Florida: /ɔ/' (0.896), and 'lawyer: /ɔj/' (0.879). The 'cot/caught' distinction had a probability of 0.507.

In the New York City and the Mid-Atlantic States dialect region, the highest probable variables were 'handkerchief: /ɪ/' (0.995), 'Monday, Friday: /e:/' (0.945), 'pajamas: /ɑ/' (0.927), 'been: /ɪ/' (0.917), 'route: /ru:t/' (0.906), 'lawyer: /ɔj/' (0.880), and 'coupon: /u:/' (0.876). This dialect had the highest probability of the 'cot/caught' distinction at 0.828.

In answer to my second research question, the results from a relatively novel statistical analysis, the latent class analysis, in this thesis confirm what the Linguistic Atlas of North American English and Katz's heat maps of the Harvard Dialect Survey have already found, albeit with some important exceptions. One of the major discrepancies between the results from the latent class analysis and the linguistic atlas is the region of the low back merger. In the latent class analysis, the North dialect has a low probability of the 'cot/caught' low back vowel distinction, whereas according to the linguistic atlas, this is a salient variable of the North dialect. Another discrepancy is that the in the West dialect, the pronunciation of 'pajamas' with the vowel /ɑ/ has a probability of 0.970 and with the vowel /æ/ at 0.000. The heat map for this variable predicts a pronunciation with /æ/ rather than /ɑ/. Also, according to the heat maps, "route" should have a higher pronunciation of "r/aʊ/t" in the West dialect. Yet, the West dialect has a higher probability of the "r/u:/t" pronunciation. Similarly, the variable "lawyer" does not correspond correctly between the Midland dialect from the latent class analysis and the heat maps. The heat maps show the Midland dialect as having both "l/ɔj/er" and "l/ɑ/yer" as possible

pronunciations for “lawyer”. However, cluster V has a high probability of the “l/ɔj/er” pronunciation at 0.879 and a low probability for the “l/ɑ/yer” pronunciation (0.163).

The discrepancies between the latent class analysis and the linguistic atlas as well as the heat maps show how different analyses can produce different results. They also suggest the power of using a multivariate approach to better understand all of the phonetic variation between dialects, by enabling the consideration of multiple variables simultaneously. Nevertheless, the findings from the latent class analysis are basically consistent with the previous classifications based on the linguistic atlas and the heat maps.

This latent class analysis has several strengths. First, it is based on a large dataset of American English dialects. A second strength is that the data take into account several linguistic variables. Another strength is the use of a multivariate analysis to statistically divide the data into six naturally occurring cluster while taking into account multiple phonetic variables simultaneously. Additionally, the database included linguistic data from each US state.

Despite these strengths, several limitations require some consideration. One limitation is that the variables between the Atlas of North American and the Harvard Dialect Survey do not completely align, making it difficult to compare the two databases. Taking this into consideration, however, it is interesting to note that even slight differences in phonetic variables affect dialect groupings. Nonetheless, even when using somewhat different variables and a different method of analysis, the results were similar enough to show that the major dialects defined by the Atlas of North America still hold with the new variables from the Harvard Dialect Survey.

An additional limitation of this study was that the data were gathered via an online survey. While surveys are convenient, cost effective, and easily accessible to multiple people,

they contain some methodological weaknesses. One is that people might not be linguistically aware of their own speech. What people think they are saying might not correctly match what they are actually saying. This being so, the response they answer on a survey about their speech may not be correct, unknown to the participant and biasing the research data.

Similarly, the survey used rhyming words to best capture the true pronunciation of certain words and sounds of the participant. However, this method is potentially faulty as it is really eliciting for the sound in the rhyming word pairs. The survey also had a question asking the participants where they were from. The participants gave their city and zip code to answer this. However, it is extremely common for people to move around from city to city and even state to state. This makes the question extremely difficult to answer as people may live in a different state that they were born in or spent most of their childhood or young adult years. This is a crucial factor that leads to the dynamic aspect of dialects and language change.

Furthermore, this survey did not elicit for information regarding socioeconomic status, urban/rural setting, or ethnicity. Research has found that these factors can influence language (Edwards, 2009). This could be a reason that only three of the clusters from the analysis matched up perfectly with the heat maps. Furthermore, other linguistic atlases or linguistic descriptions describe dialects by including ethnic and socio-economic groupings as well as geographical regions (Schneider, 2008).

In conclusion, in the context of the limitations associated with this study, this analysis contributes dialectal understanding of American English because it shows how a new statistical technique can be used in dialectology. Specifically, it shows that a latent class analysis can be used in dialect studies to separate out dialect data into clusters, including the probabilities of linguistic features occurring in each dialect. Furthermore the results from the latent class analysis

also show that the results from the Harvard Dialect Survey generally parallel the findings of the Linguistic Atlas of North American English, providing support for six basic dialects of American English. This thesis also contributes to our understanding of language variation by showing how the probabilities of individual features changes throughout each dialect.

APPENDIX

I. Harvard Dialect Survey Questions

1. *au*nt
   a. /ɑ/ as in "ah" (9.62%)
   b. /æ/ as in "ant" (75.15%)
   c. /ɒ/ as in "caught" (2.77%)
   d. I have the same vowel in "ah", "caught", and "aunt" (2.52%)
   e. I pronounce it the same as "ain't" (0.58%)
   f. I use /ɑ/ɒ/ when referring to the general concept of an aunt, but /æ/ when referring to a specific person by name. (6.64%)
   g. I use /æ/ when referring to the general concept of an aunt, but /ɑ/ɒ/ when referring to a specific person by name. (1.84%)
   h. other (0.88%)
   (11713 respondents)

2. b*ee*n
   a. /ɪ/ as in "sit" (64.82%)
   b. /i:/ as in "see" (3.59%)
   c. /ɛ/ as in "set" (28.60%)
   d. other (2.99%)
   (11609 respondents)

3. the first vowel in "B*o*wie knife"
   a. /o:/ as in "Bo" (70.58%)
   b. /u:/ as in "boo" (19.27%)
   c. I have seen this word in print, but have no idea how to pronounce it(5.42%)
   d. I have never seen or heard this word (3.70%)
   e. other (1.03%)
   (11636 respondents)

4. caramel
   a. with 2 syllables ("car-ml") (38.02%)
   b. with 3 syllables ("carra-mel") (37.66%)
   c. I use both interchangeably (17.26%)
   d. I have both forms, but the two have different meanings (please state how in the comments box) (3.77%)
   e. other (3.28%)
   (11609 respondents)

5. the vowel in the second syllable of "caul*i*flower"
   a. /i:/ as in "see" (31.52%)
   b. /ɪ/ as in "sit" (63.97%)

c. other (4.51%)
(11575 respondents)

7. c*ou*pon
a. with /u:/ as in "coop" ("coopon") (66.86%)
b. with /ju:/ as in "cute" ("cyoopon") (31.31%)
c. other (1.83%)
(11571 respondents)

8. Cr*ai*g (the name)
a. /ɛ/ as in "set" (28.00%)
b. /e:/ as in "say" (40.17%)
c. I say something in between the vowels in "set" and "say", but closer to the one in "say" (17.48%)
d. I say something in between the vowels in "set" and "say", but closer to the one in "set" (13.46%)
e. other (0.90%)
(11519 respondents)

9. cr*ayo*n
a. /æ/ as in "man" (1 syllable, "cran") (14.13%)
b. /ejɑ/ (2 syllables, "cray-ahn") (48.64%)
c. /ejɒ/ (2 syllables, "cray-awn", where the second syllable rhymes with "dawn") (34.53%)
d. /aw/ (I pronounce this the same as "crown") (1.46%)
e. other (1.24%)
(11514 respondents)

10. cr*ee*k (a small body of running water)
a. /i:/ as in "see" (88.57%)
b. /ɪ/ as in "sit" (3.85%)
c. I use both interchangeably (5.13%)
d. I don't know how to pronounce this word (0.04%)
e. I use both, but they mean two different things (please state how they differ in the comments box) (2.05%)
f. other (0.36%)
(11517 respondents)

11. the first vowel in "Fl*o*rida"
a. /o:/ as in "flow" ("flow-ri-da") (4.95%)
b. /ɑ/ as in "ah" ("flah-ri-da") (11.37%)
c. /ɒ/ as in "saw" ("flaw-ri-da") (7.09%)
d. /ɔ/ as in "sore" ("flore-i-da") (73.38%)
e. other (3.20%)
(11451 respondents)

12. fl*ou*rish
a. /ə/ as in "b*i*rd" ("flurr-ish") (62.23%)
b. /ɔ/ as in "sore" ("flore-ish") (23.07%)
c. /ʌ/ as in "sun" ("fluh-rish") (10.18%)
d. other (including if you use one pronunciation for the verb and a different pronunciation for the noun) (4.52%)
(11429 respondents)

13. the last vowel in "handkerch*ie*f"
a. /i:/ as in "see" (19.96%)
b. /ɪ/ as in "sit" (78.23%)
c. other (1.81%)
(11400 respondents)

14. l*aw*yer
a. with /ɔj/ as in "boy" ("loyer") (72.84%)
b. with /ɒ/ as in "saw" ("law-yer") (21.96%)
c. I use both interchangeably (4.86%)
d. other (0.34%)
(11421 respondents)

16. m*ayo*nnaise
a. with /æ/ as in "man" (2 syllables--"man-aze") (41.65%)
b. with /ejə/ (3 syllables--"may-uh-naze") (45.83%)
c. I use both interchangeably (8.81%)
d. other (3.71%)
(11372 respondents)

17. the first vowel in "m*i*racle"
a. /i:/ as in "near" (26.21%)
b. /ɪ/ as in "knit" (52.13%)
c. /ɛ/ as in "net" (2.35%)
d. I say something in between /ɪ/ and /ɛ/ (15.38%)
e. other (3.94%)
(11284 respondents)

19. the final vowel in "Mond*ay*," "Frid*ay*," etc.
a. /e:/ as in "say" (86.78%)
b. /i:/ as in "see" (4.69%)
c. I use /e:/ with the words in isolation, but /i:/ in compounds (such as "Sunday school") (6.12%)
d. other (e.g. do you use one vowel in some day names, and another in the other names?) (2.40%)
(11316 respondents)

20. the second vowel in "paj*ɑ*mas"
    a. /æ/ as in "jam" (45.92%)
    b. /ɑ/ as in "father" (51.86%)
    c. other (2.23%)
    (11277 respondents)

21. pecan
    a. /pi:kæn/ with stress on the first syllable ("PEE-can") (17.03%)
    b. /pi:kæn/ with stress on the second syllable ("pee-CAN") (9.02%)
    c. /pi:kɑn/ with stress on the first syllable ("PEE-Kahn") (13.19%)
    d. /pi:kɑn/ with stress on the second syllable ("pee-KAHN") (28.60%)
    e. /pɪkæn/ ("pick Ann") (1.48%)
    f. /pɪkɑn/ ("pick Ahn") (20.92%)
    g. I pronounce it differently when it's alone than when it's in a compound like "pecan pie" (please state how you pronounce the two variants in the comments box) (6.24%)
    h. other (3.51%)
    (11213 respondents)

22. poem
    a. one syllable (32.39%)
    b. two syllables (67.61%)
    (11235 respondents)

23. r*ea*lly
    a. /i:/ as in "see" ("reely") (52.54%)
    b. /ɪ/ as in "sit" ("rilly") (26.28%)
    c. /iə/ ("ree-l-y") (8.21%)
    d. other (including if you use two or more of these interchangeably)(12.97%)
    (11175 respondents)

24. realtor (a real estate agent)
    a. 2 syllables ("reel-ter") (44.21%)
    b. 3 syllables (real/ə/tor, in other words "reel-uh-ter") (32.21%)
    c. 3 syllables (ree-l-ter) (19.70%)
    d. I don't use this word; I use "estate agent" (1.09%)
    e. other (2.79%)
    (11148 respondents)

26. route (as in, "the route from one place to another")
    a. rhymes with "hoot" (29.99%)
    b. rhymes with "out" (19.72%)
    c. I can pronounce it either way interchangeably (30.42%)
    d. I say it like "hoot" for the noun and like "out" for the verb. (15.97%)
    e. I say it like "out" for the noun and like "hoot" for the verb. (2.50%)

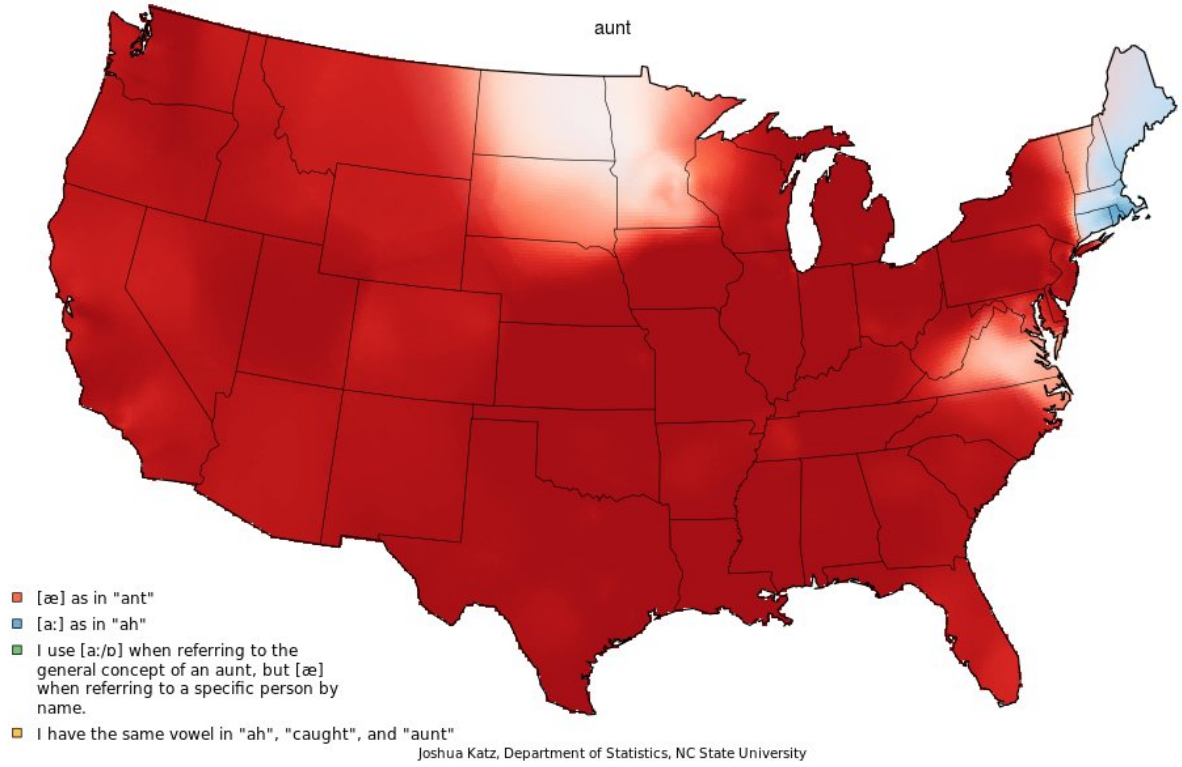
f. other (1.40%)
(11137 respondents)

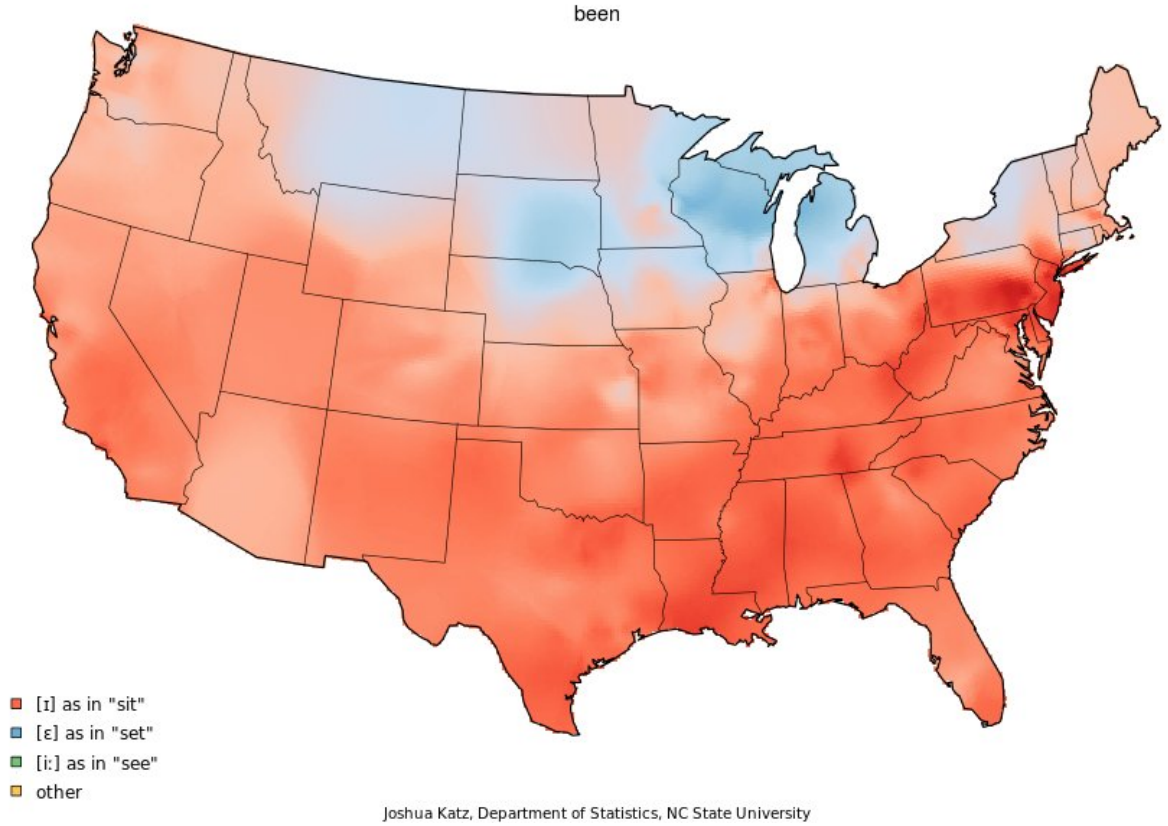
27. the first vowel in "s*y*rup"
a. /i/ "sear-up" (13.43%)
b. /ɪ/ "sih-rup" (34.08%)
c. /ə/ as in "sir" (49.89%)
d. other (2.60%)
(11107 respondents)

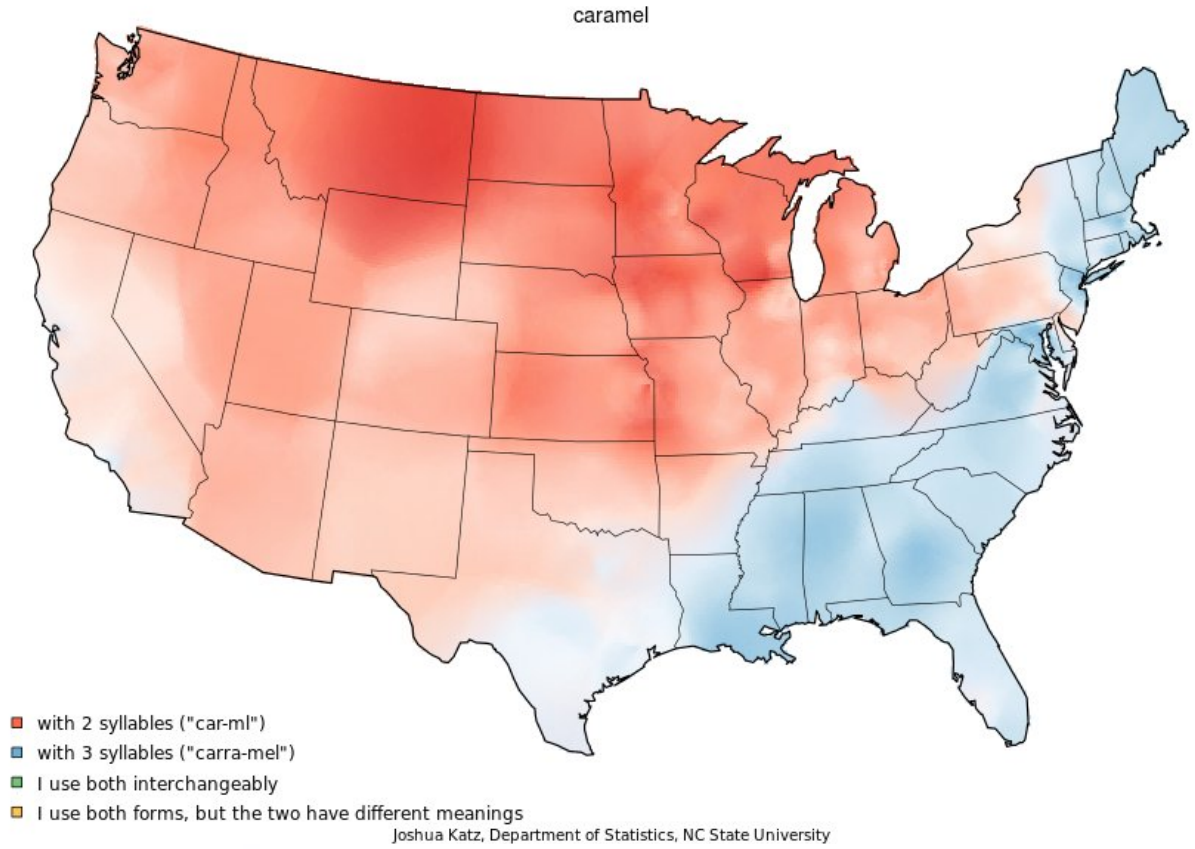
28. Do you pronounce "cot" and "caught" the same?
a. different (60.93%)
b. same (39.07%)
(11050 respondents)

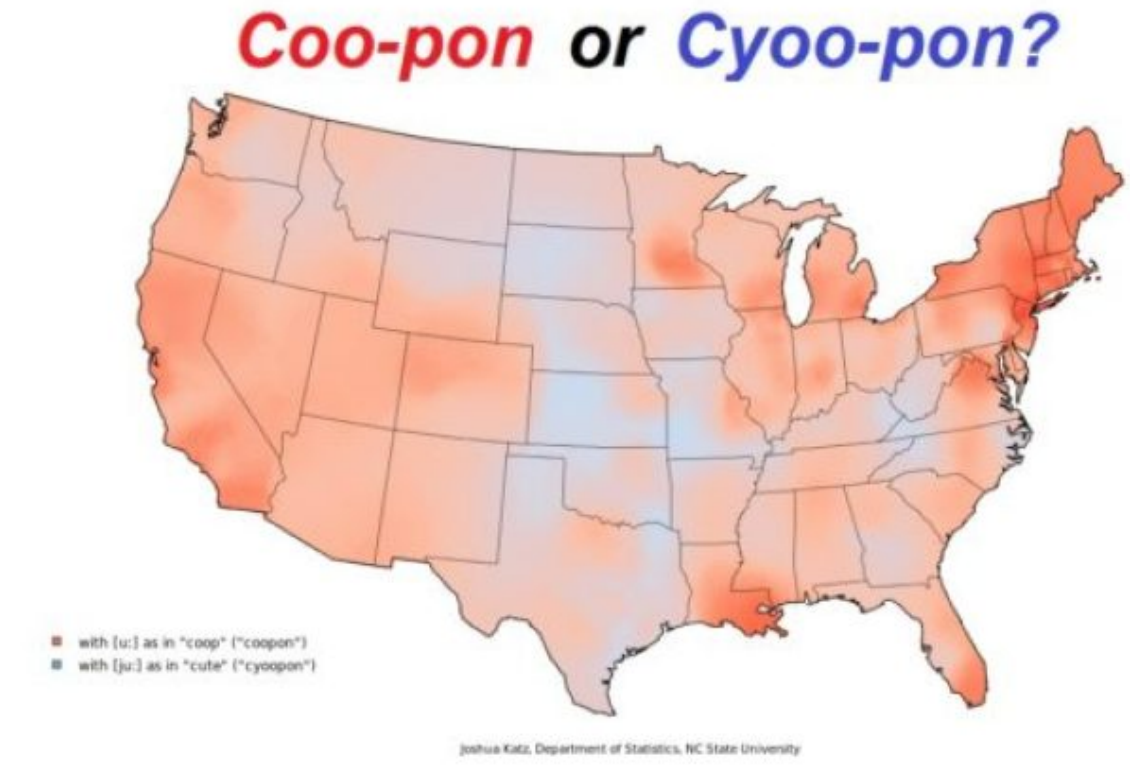
108. What vowel do you use in b*a*g?
a. /æ/ as in "sat" (88.62%)
b. /ɛ/ as in "set" (0.56%)
c. /e:/ as in "say" (8.42%)
d. other (2.40%)
(10632 respondents)

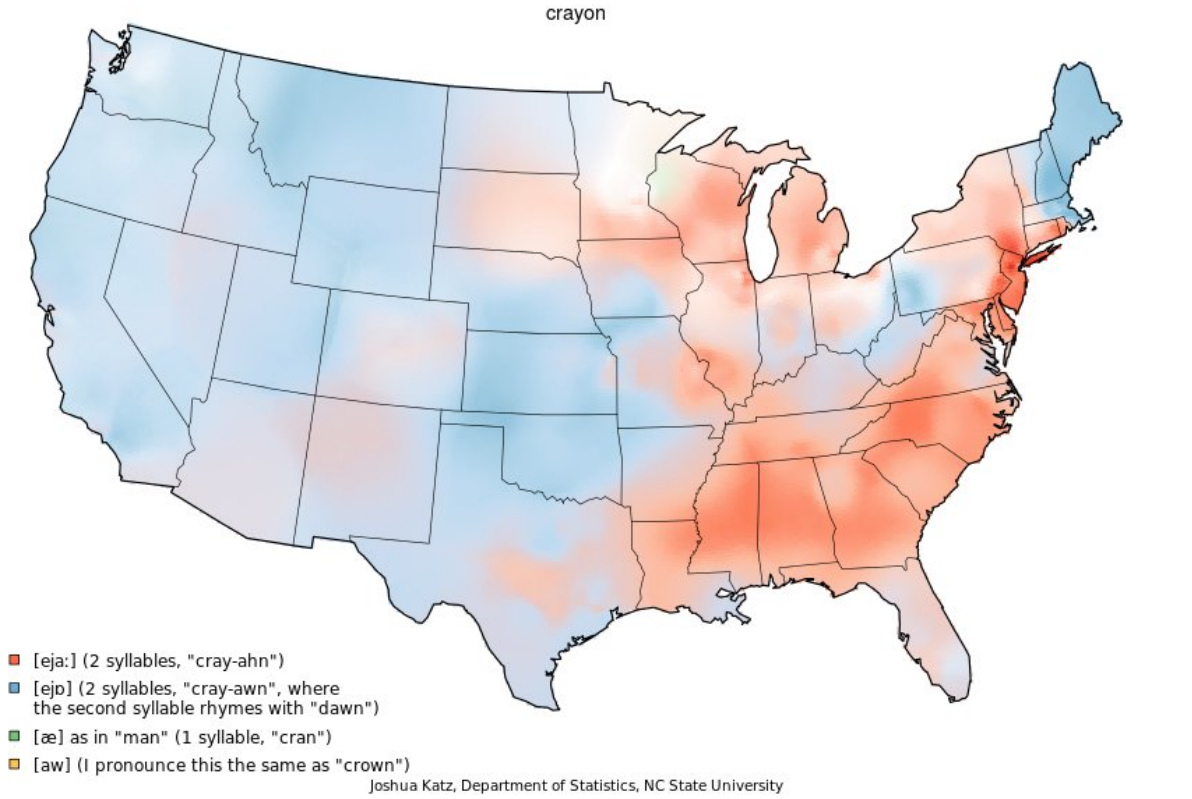
KATZ HEAT MAPS

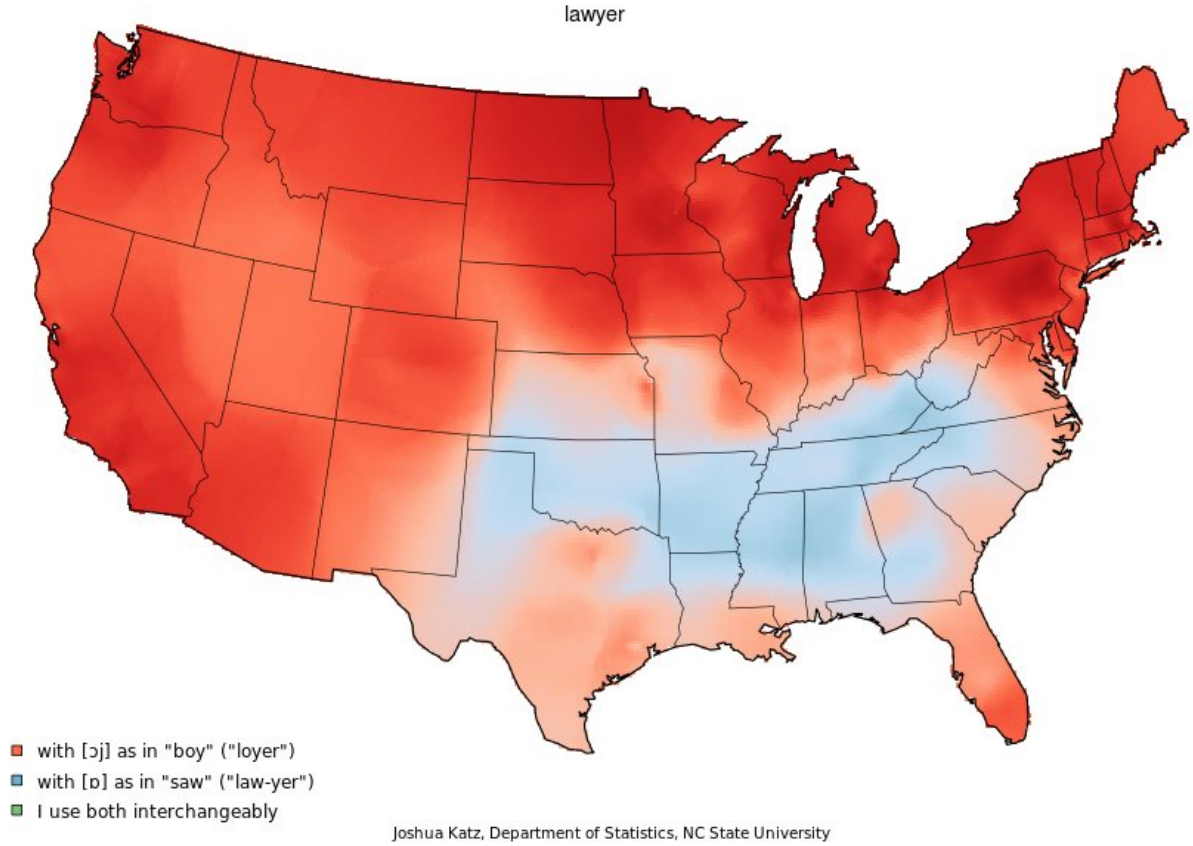


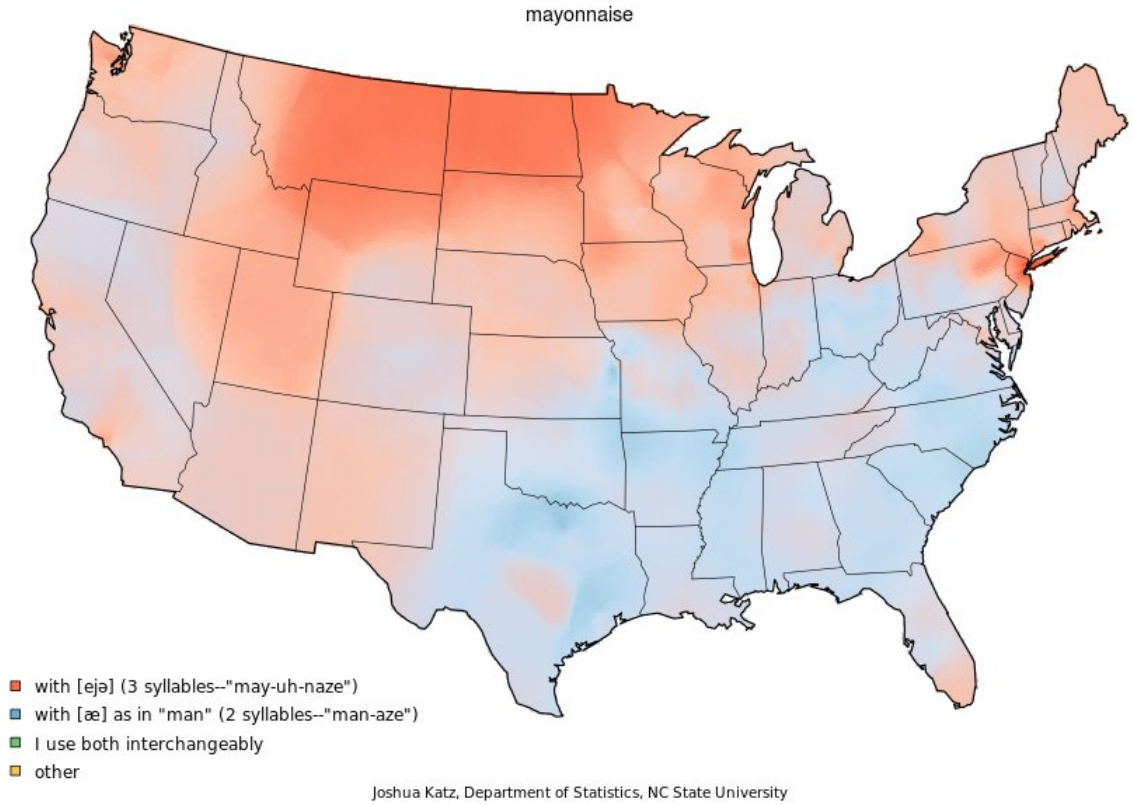


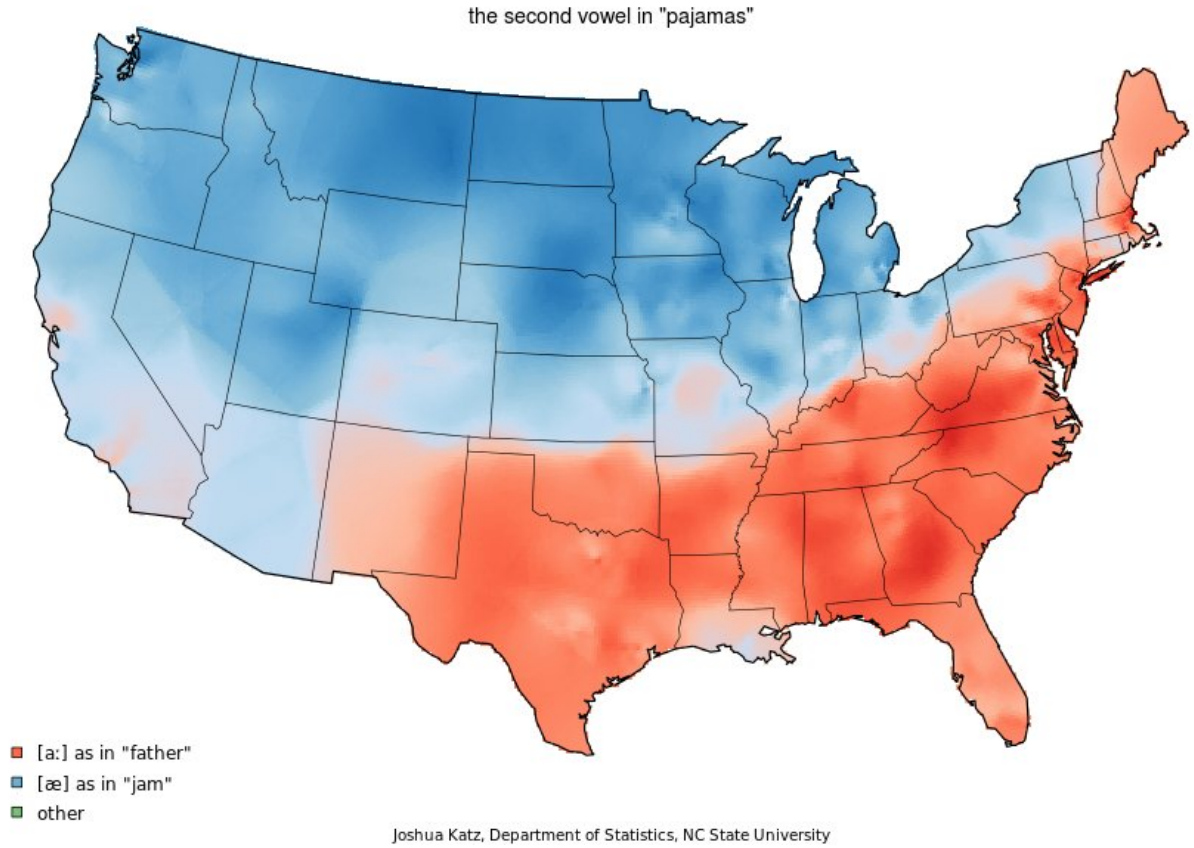

caramel
with 2 syllables ("car-ml")
with 3 syllables ("carra-mel")
I use both interchangeably
I use both forms, but the two have different meanings
Joshua Katz, Department of Statistics, NC State University

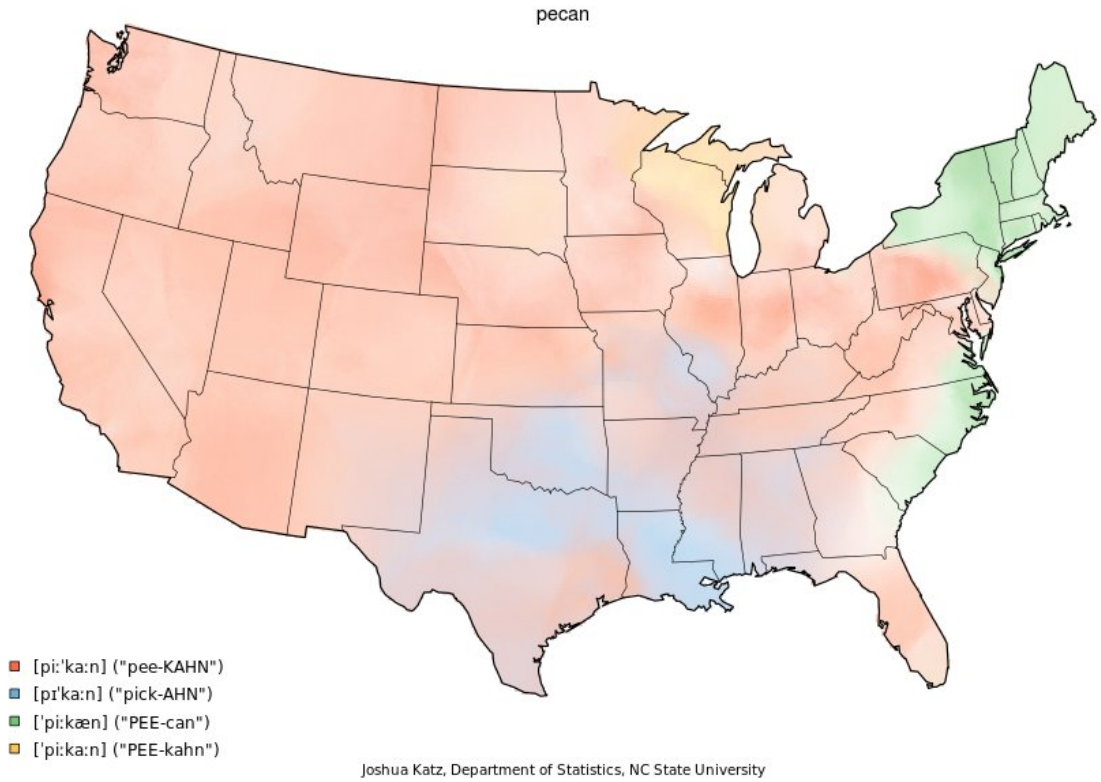

Coo-pon or Cyoo-pon?
with [u:] as in "coop" ("coopon")
with [ju:] as in "cute" ("cyoopon")
Joshua Katz, Department of Statistics, NC State University

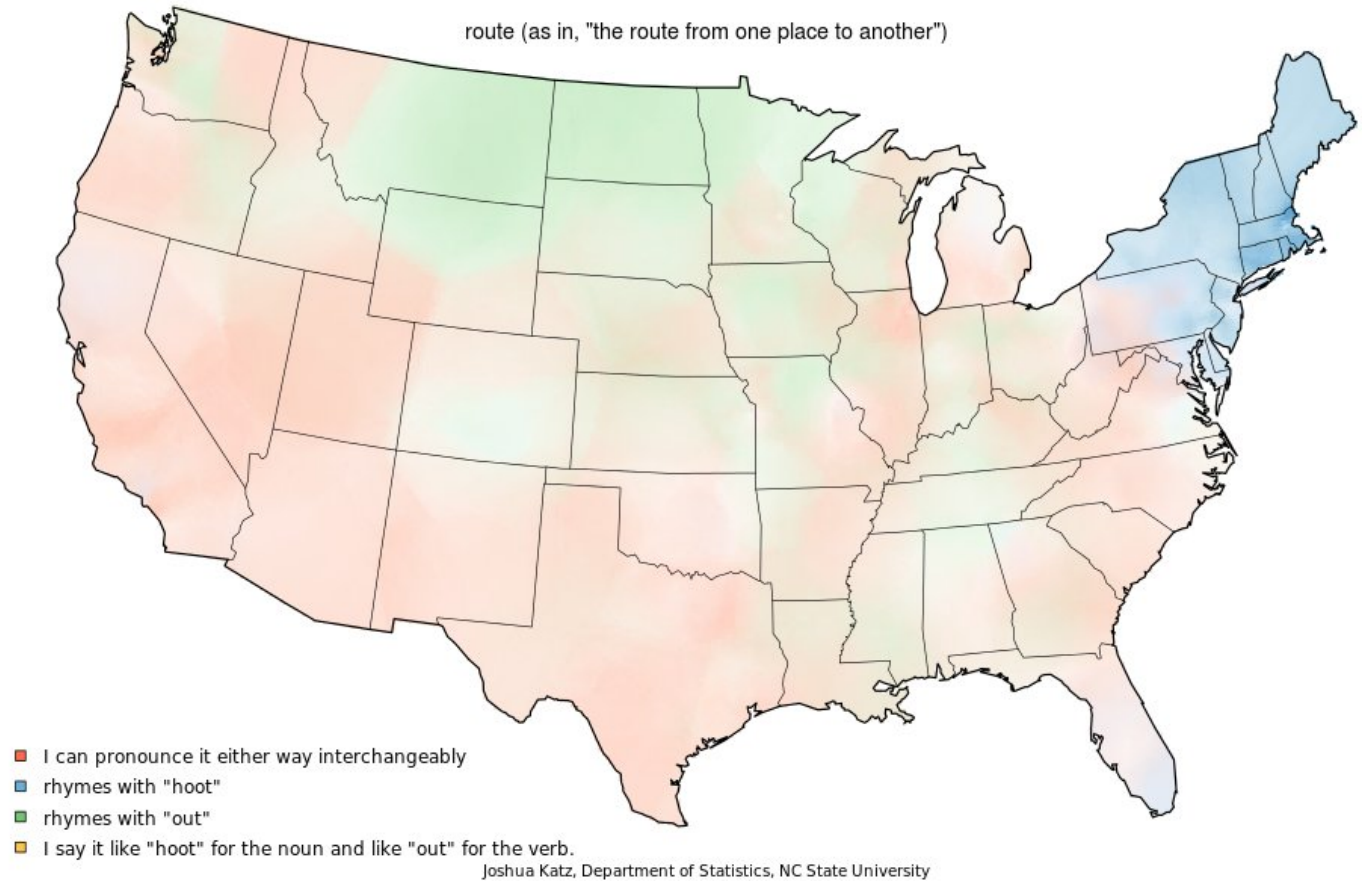

crayon
[eja:] (2 syllables, "cray-ahn")
[ejɒ] (2 syllables, "cray-awn", where the second syllable rhymes with "dawn")
[æ] as in "man" (1 syllable, "cran")
[aw] (I pronounce this the same as "crown")
Joshua Katz, Department of Statistics, NC State University

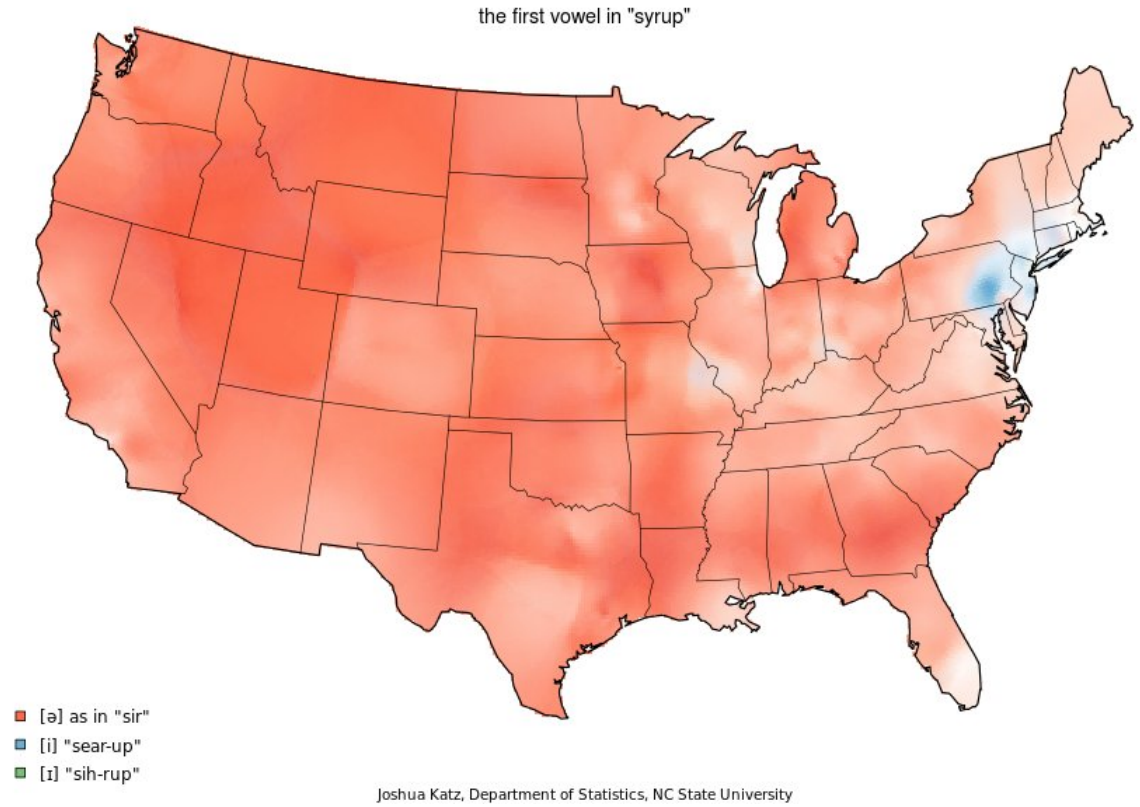

lawyer
with [ɔj] as in "boy" ("loyer")
with [ɒ] as in "saw" ("law-yer")
I use both interchangeably
Joshua Katz, Department of Statistics, NC State University

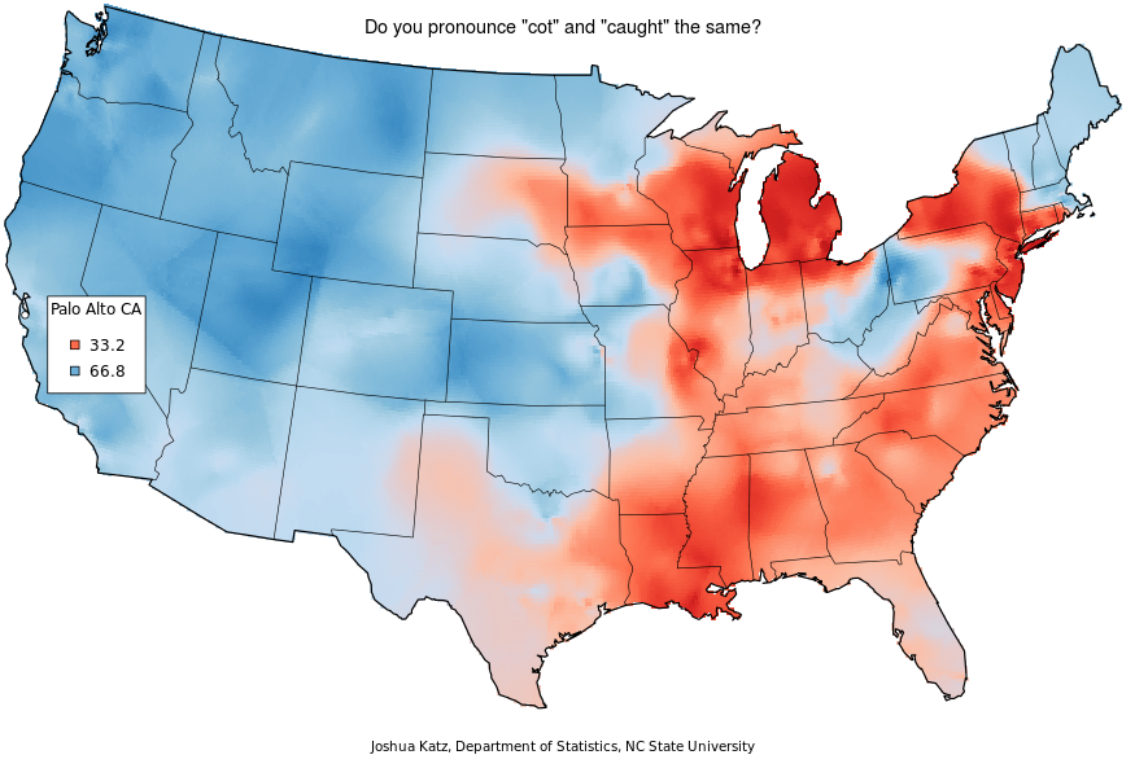

mayonnaise
with [ejə] (3 syllables--"may-uh-naze")
with [æ] as in "man" (2 syllables--"man-aze")
I use both interchangeably
other
Joshua Katz, Department of Statistics, NC State University


the second vowel in "pajamas"
[a:] as in "father"
[æ] as in "jam"
other
Joshua Katz, Department of Statistics, NC State University


pecan
[piː'kaːn] ("pee-KAHN")
[pɪ'kaːn] ("pick-AHN")
['piːkæn] ("PEE-can")
['piːkaːn] ("PEE-kahn")
Joshua Katz, Department of Statistics, NC State University


route (as in, "the route from one place to another")
I can pronounce it either way interchangeably
rhymes with "hoot"
rhymes with "out"
I say it like "hoot" for the noun and like "out" for the verb.
Joshua Katz, Department of Statistics, NC State University


the first vowel in "syrup"
[ə] as in "sir"
[i] "sear-up"
[ɪ] "sih-rup"
Joshua Katz, Department of Statistics, NC State University


Do you pronounce "cot" and "caught" the same?
Palo Alto CA
33.2
66.8
Joshua Katz, Department of Statistics, NC State University

REFERENCES

Abdi, H. (2003). Multivariate analysis. Encyclopedia for research methods for the social sciences. Thousand Oaks: Sage, 699-702.

Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS TwoStep Cluster-a first evaluation.

Banuazizi, A., & Lipson, M. (1998). The tensing of/ae/before/l/: An anomalous case for short-a rules of white Philadelphia speech. Papers in sociolinguistics: NWAVE-26 à l'Université Laval, 41-52.

Biber, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. Linguistics, 23(2), 337-360.

Byrne, Richard. (2013, June 17). 128 Maps of Regional Dialect Differences. *Free Technology for Teachers*. Retrieved from http://www.freetech4teachers.com/2013/06/128-maps-of-regional-dialect-differences.html#.WSOlKmjyvIU.

Coloma, G. (2011). The Socio-Economic Significance of Four Phonetic Characteristics in North American English. Serie Documentos de Trabajo Document, (459).

Conduct and Interpret a Cluster Analysis. (2017) In *Statistical Solutions: Advancement Through Clarity,* Retrieved from http://www.statisticssolutions.com/cluster-analysis-2/

Eddington, D., & Channer, C. (2010). American English has goʔ a loʔ of glottal stops: Social diffusion and linguistic motivation. American speech,85(3), 338-351.

Edwards, J. (2009). Language and identity. Blackwell Publishing Ltd.

Ferguson, C. A. (1972). Short a'in Philadelphia English. *Studies in Linguistics in Honor of George L. Trager*, Mouton.

Ferguson, Charles. (1994). Sociolinguistic Perspectives on Register. D. Biber and E. Finegan (Eds.). New York, NY: Oxford University Press.

Grieve, J. (2012). A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. Corpus Linguistics and Linguistic Theory, 8(1), 39-72.

Hagiwara, R. (1997). Dialect variation and formant frequency: The American English vowels revisited. *The Journal of the Acoustical Society of America*, 102(1), 655-658.

Hamblin, J. (2013, June 6). Pecan, Caramel, Crawfish: Food Dialect Maps. *The Atlantic*. Retrieved from https://www.theatlantic.com/health/archive/2013/06/pecan-caramel-crawfish-food-dialect-maps/276603/.

Hickey, Walter. (2013, June 5). 22 Maps That Show How Americans Speak English Differently From One Another. *Business Insider*. Retrieved from http://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6/#the-pronunciation-of-caramel-starts-disregarding-vowels-once-you-go-west-of-the-ohio-river-1.

Hyvönen, S., Leino, A., & Salmenkivi, M. (2007). Multivariate Analysis of Finnish Dialect Data—An Overview of Lexical Variation. Literary and Linguistic Computing, 22(3), 271-290.

Katz, J., Andrews, W., and Bluth, E. (2013) How Y'all, Youse and You Guys Talk. *The New York Times*. Retrieved from http://www.nytimes.com/interactive/2013/12/20/sunday-review/dialect-quiz-map.html?_r=0.

Katz, Joshua. (2013). *Beyond "Soda, Pop, or Coke: Regional Dialect Variation in the Continental US*. Retrieved from http://www4.ncsu.edu/~jakatz2/project-dialect.html.

Kleinman, Alexis. (2013). These Dialect Maps Showing The Variety Of American English Have Set The Internet On Fire. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/2013/06/06/dialect-maps_n_3395819.html.

Kupzyk, K. A. (2011). Introduction to Mixture Modeling. Retrieved from http://r2ed.unl.edu/presentations/2011/RMS/012111_Kupzyk/012111_Kupzyk.pdf

Labov, W. (1963). The social motivation of a sound change. Word, 19(3), 273-309.

Labov, W. (1989). Exact description of the speech community: Short A in Philadelphia. Language Change and Variation, 1-57.

Labov, W., Ash S., & Boberg C. (2005a). [Graph illustrations of phonetics, phonology, and sound changes]. *The Atlas of North American English: Phonetics, Phonology and Sound Change.* Retrieved from http://www.atlas.mouton-content.com/.

Labov, W., Ash S., & Boberg, C. (2005b). The Atlas of North American English: Phonetics, Phonology and Sound Change. Berlin/Boston, DE: De Gruyter Mouton. Retrieved from http://www.ebrary.com.erl.lib.byu.edu

Lakoff, R. (1990). Why can't a woman be less like a man?'. *Talking Power: The Politics of Language*. San Francisco, CA: Basic Books.

Metcalf, A. A. (2000). How we talk: American regional English today. Houghton Mifflin Harcourt.

Roberts, J. (1997). Acquisition of variable rules: a study of (-t, d) deletion in preschool children. Journal of Child Language, 24(02), 351-372.

Roberts, J., & Labov, W. (1995). Learning to talk Philadelphian: Acquisition of short a by preschool children. Language Variation and Change, 7(01), 101-112.

Schneider, E. (Ed.) (2008). *2 The Americas and the Caribbean*. Berlin, Boston: De Gruyter Mouton.

Skendi, S. (1975). Language as a Factor of National Identity in the Balkans of the Nineteenth Century. Proceedings of the American Philosophical Society, 119(2), 186-189.

Vaux, B., & Golder, S. (2003). The Harvard dialect survey. Cambridge, MA: Harvard University Linguistics Department.

Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. Annu. Rev. Linguist., 1(1), 243-264.

Xiao, R. (2009). Multidimensional analysis and the study of world Englishes.World Englishes, 28(4), 421-450.