**Statistical Design Consulting**

**SEMESTER REPORT**

**Fall 2024**

**Client:** Leah Nodar  **File Number:** 24-050

**Department:** Linguistics **Major Prof:** Dr. Felicia Roberts

**Consultant:** Sumeeth Guda **Initial Meeting Date:** 07/10/24

**Meeting Attendees:** Leah Nodar, Sumeeth Guda, Dr. Felicia Roberts, Dr. Bruce Craig

**Statement of Problem:**

To determine whether groups of speakers from Mobile and Africatown, Alabama have significantly different speech patterns, and investigate how the dialects evolved over time using seven linguistic features

**Goal of This Project:** PhD dissertation

**Background:**

The client is a PhD student in the linguistics department who is exploring whether two groups of people from the early 20th century, who had very different English dialects, continue to have these differences in the 1970s and 1990s, or if their dialects have converged.

The two groups are the people of Africatown, AL and those of Mobile, AL. According to the client, this is linguistically interesting for two reasons. First, because Africatown has a unique dialect that has not been studied before, and second, because Africatown has documentation of a town founder's earliest stage of the dialect in the 1920s, so following this across later time periods may tell us something new about how dialects develop overall, with implications for how other languages change.

The client will not be conducting an experiment, but rather perform a social/historical analysis. The client’s data are constrained to the audio available to the client from the two time periods mentioned, with a small sample within the groups. The client has two time periods they are working under (1970s, 1990s), and they have three groups = {Group A: The descendants of the original Africatown founders, Group B: People who were born and raised in Africatown but are not direct descendants, Group C: people from Mobile, AL}.

The distribution of the groups is:

1970s :

* Group A: 2 people
* Group B: 2 people
* Group C: 7 people

1990s:

* Group A: 3 people
* Group B: 4 people
* Group C: 11 people

The way the client is currently gathering the data is by listening to each of the people’s audio and marking up / flagging seven linguistic variables (pronunciations or grammatical features). These seven variables were chosen based on previous research that analyzed the documentation of speech from the 1920s of a founder of Africatown.

The client indicated the features she was looking out for in her attached document “SevenFeatures”.

As an example, one feature is the pronunciation of "th" sounds as "d" (as in "dey, dese, dose" vs "they, these, those"). For each person, the client followed linguistic criteria to determine ~60 potential environments (places where this sound could occur) and then checked the audio and spectrogram for each of those environments to determine whether the person actually pronounced it with "th" or "d" (or neither).

The other features are similar: Based on linguistic criteria, determination of the environments in each person's speech where a feature could occur, followed by a check for whether it does occur. Generally, there are at least 50-100 potential environments per feature.

The client’s goal is to determine whether as a whole this collection of seven features indicates that the groups of speakers have significantly different speech patterns, with a broader goal of exploring how dialects evolve over time.

**Progress During Current Semester:**

This past semester Leah finished collecting her transcribed data from the interviews of the participants. After she put them in a dataset, Leah fit 2 different logistic regression models for each of the data sets with the response variable being the generation {70s, 90s}. She did this for 2 of her linguistic features. For the most part this semester her main challenge was understanding the steps needed to do the fitting, as well as how to evaluate the fit. Sumeeth explained how to use residual deviance and compare it against the degrees of freedom of the model to evaluate the fit. Afterwards to compare the two generations to establish there was difference between them, Sumeeth suggested that she look at the Wald test, bootstrapping, and anova chisq tests as methods to compare the differences and determine that the models are different. There are some challenges since her data is limited in size, and she can’t collect any more data due to funding and the fact that her data is historical. Hence she will have to continue into the spring semester.

**Current Status: Continuing.**