Client 24-050

Meeting time: 07/10, Wednesday, 10:30 AM

Current stage: Presently Collecting Data

Expected completion date: 10/31/2024

Supported by a grant or contract: No

Background:

The client is a PhD student in the linguistics department who is exploring whether two groups of people who grew up in the early 20th century spoke very different dialects of English continue to speak differently 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 2 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 language change.

The client will not be conducting an experiment, but rather a social/historical analysis. The client’s data is constrained to the audio available to the client from the two time periods mentioned, with a small population within the groups. The client has 2 time periods they are working under (1970s, 1990s), and they have 3 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 variables 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 variable.

When the data markup is complete, the client’s desired outcome is to determine whether as a whole this collection of seven features indicates that the groups of speakers have significantly different speech patterns.

Research questions from client:

1. In the 1970s, do descendants of Africatown founders (Group A) pattern significantly differently in speech than Mobilians (Group C) with respect to the collection of seven linguistic features under consideration?
2. In the 1970s, do people raised in Africatown (Groups A and B together) pattern significantly differently in speech than Mobilians (Group C) with respect to the collection of seven linguistic features under consideration?
3. In the 1990s, do descendants of Africatown founders (Group A) pattern significantly differently in speech than Mobilians (Group C) with respect to the collection of seven linguistic features under consideration?
4. In the 1990s, do people raised in Africatown (Groups A and B together) pattern significantly differently in speech than Mobilians (Group C) with respect to the collection of seven linguistic features under consideration?

Statistical issues

1. What is the most appropriate statistical process for testing differences between groups containing speakers containing linguistic variables?
2. Are there additional steps client can take to increase their confidence in the results despite the small number of speakers?

Discussion points:

1. Are the people from Mobile the control group in this study?
2. Does environment mean group labeling?
3. If the client’s main mission is to just see if the groups are different, Anova or Tukey HSD could be used to just show that the groups are different.
4. With such a small sample size, ultimately there might be lack of fit issues if regression is used.
5. Does averaging out each subgroup matter for this experiment?
6. How does the dataset look like?