MANCHESTER 1824
The University of Manchester

Lancaster University: Analysing Big Data to study evolution of UK Dialects

Enhui Cheng & Chloe Gornall, BA (Hon) Linguistics

Introduction

We completed a 7-week Q-Step internship with respect to sociolinguistics research in the University of Lancaster. It involved processing and analysing data from the English Dialects App corpus, which contained audio recordings from more than 3700 speakers from across the UK. The data was used to examine speech variation across the UK and the results will inform research on speaker profiling and speaker identification.

Objectives

The ultimate goal of the research was to explore how phonetic features varied across the UK and how they evolved over time. However, the aim of our internship was to discover any factors that led to the variation of Northern English dialects. We were expected to analyse the data to inspire the supervisor with more new ideas for the research. In our project, rather than dialects from across UK, we concentrated on Northern English dialects. To be more specific, we only analysed the data among the speakers from 7 northern cities - Liverpool, Manchester, Leeds, York, Sheffield, Newcastle Upon Tyne and Hull. In addition, instead of all the speech sounds, we mainly focused on the articulation of vowels. In order to find any potential breakthrough points, we took different sets of vowels as research objects respectively. First set was the vowels in "trap"("ae"), "bath"("aa"), "strut"("ah"), and "foot" ("uh"). Second set was vowels in "fleece", "goose" and "thought". Third set was monophthongs and diphthongs. Fourth set was all vowels.

Methodology

The original dataset was generated from English Dialect App corpus, which contained audio recordings and personal details from more than 3,700 speakers from across the UK. All the speakers using the App were required to read the story "The Boy Who Cried Wolf", which consisted of ten sentences. The recordings of each sentence were saved separately for the sake of convenience for force alignment and research objects selection. Personal details included the speaker's gender, age, education status, address, ethnicity, etc. In our project, we only looked into the influence of age, gender, city, mobility (frequency of residence change in ten years) on the articulation of vowels. In terms of city, we only focused on Liverpool, Manchester, Leeds, Manchester, York, Sheffield, Newcastle Upon Tyne and Hull. Therefore, we got rid of all the other factors' columns and all the speakers from cities other 7 cities indicated above. We selected 10 "perfect" speakers from each city, which meant each speaker must have recordings with good quality of all the ten sentences. The data of all the spare speakers were chucked out using Excel. To help the supervisor save time, we also noted down some key phonetic features of each city through literature review.

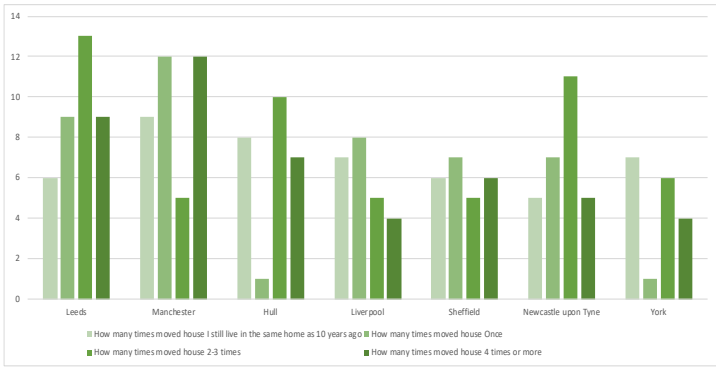

Figure 1. Mobility distribution of participants in each studied city

Key Skills

Praat:

- Transcription realignment: Transcriptions of all the recordings were generated by force alignment beforehand. However, the annotation of each phoneme didn't align perfectly with its content, so we had to listen to each recording and adjusted the alignment. Also, reannotation was required, as force alignment could be wrong sometimes..
- Formant extraction: Run Praat script to extract midpoint and dynamic formant of vowels. Only F1 and F2 needed to be measured, because they were sufficient to map the vowel on the vowel chart.

R Studio:

- Making boxplots to compare the variance of formants value in different cities and investigate the effects of geographic factors on articulation
- Run statistical tests: Running two-way ANOVA to test whether city and mobility have significant effect on F1 and F2 of vowels "aa", "ae", "ah", "uh". Running t-test to test whether mobility is a significant factor on the articulation variation in Leeds and Hull.

Microsoft Office Excel:

- Using calculation and manipulation formulae such as "= countif" to count the amount of speakers with respect to different categories and help to chuck out spare speakers.
- Making bar charts to visualise demographics.
- Making vowel plots to map vowels onto it based on its formant value.

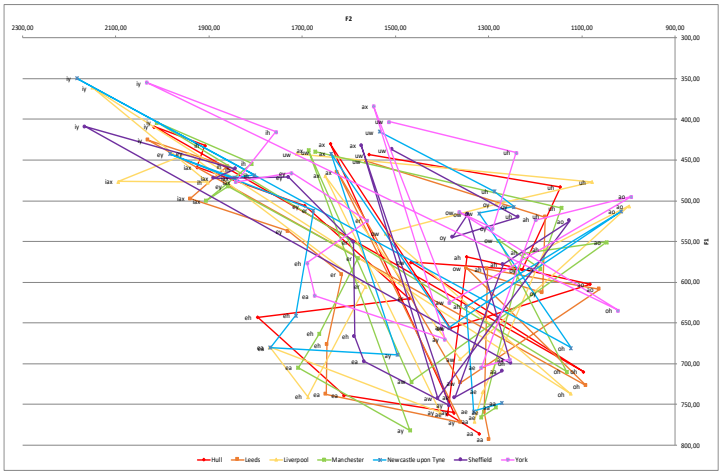

Figure 2. Findings on the average vowel space of participants in each studied city

Conclusion

In summary, our role at Lancaster was to assist the researcher in phonemic alignment corrections, as well as various data and statistical analysis. The project was not finished in the time that we were there, so we do not know the complete findings of region accent differences. However, there were some potential breakthroughs in Hull and Leeds, and some findings are shown in Figure 2. Overall, Q-step was a great invaluable experience that developed our statistical, linguistic, and social skills.

Produced at the Graphic Support Workshop, Media Services, The University of Manchester

Lancaster University

Contact
enhui.cheng@student.manchester.ac.uk & chloe.gornall@student.manchester.ac.uk

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE