Assessing research bias against English varieties: a systematic review

Caitlin Halfacre[2] and Liam Keeble[1]

[1]Henry Wellcome Building, Medical School, Newcastle Upon Tyne, NE2 4HH, United Kingdom
[2]Percy Building, School of English Literature, Language and Linguistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

1 Introduction

The literature on variation and change in English is wide-ranging but tends to be clustered around particular varieties (Trudgill & Watts 2002). Even Wells' (Wells 1982) volume on Accents of the British Isles only has one chapter on the entirety of the North of England, with specific sections on only Merseyside and Tyneside English. This is part of a larger ongoing situation where there is greater academic understanding of some varieties, and indeed some languages, due to a larger body of research existing. We would like to investigate factors which affect whether a variety is more likely to be studied, particularly focusing on whether the presence of a university Linguistics department nearby increases the research output on a particular variety, amongst other factors.

This paper sets out a novel methodology aiming to understand the current state of linguistic study across varieties of English spoken in England. We present research questions relating the location of English dialects and the extent to which they are studied in relation to centres of linguistic research, and outline methods proposed to answer them using bibliographic data, basic systematic review protocols and map data in R.

The final aim of the project is to answer the below questions. However, here we focus on outlining the approach for questions 1 and 2:

1. Are English varieties that are geographically distant from linguistics university departments more likely to be understudied?

2. Does the presence of a locally focused corpus increase the research output on a particular variety?

3. Are English varieties associated with higher social/income status lacking in research articles?

4. Is more research conducted on varieties of English associated with suburban, metropolitan, or rural areas?

2 Methods

2.1 Data extraction

Wikipedia will be used to categorise accents. If accents were defined by academic sources, there is a risk that under-studied accents would be missing from the dataset. Since we are trying to identify gaps in the research/academic literature, it becomes important to take a different approach to defining the varieties. Wikipedia is a community established encyclopedia, which provides us with an opportunity to use popular rather than academic definitions. (This could perhaps be viewed as a folk categorisation of varieties of English, and thus this study could be viewed as assessing how close linguistic research fully describes public opinion of the existence of certain varieties of English). The geographical area associated with English varieties will be ascertained from their Wikipedia entries also.

Proximity will be measured using Google maps, and data will be gathered using the 'mapdist' function from the ggmap r package (Kahle & Wickham 2013, R Core Team 2018). Proximity from the geographical area of the English variety to the nearest university, the nearest university with a Linguistics or English Language degree, the nearest sociolinguistics/language variation lab or research group, and the nearest linguistics department will all be measured and included in the dataset as separate variables. Information on the existence of research labs, linguistics departments and degrees will be found on university websites. Whether or not the variety has a corpus (ascertained from web searches), is typically associated with a metropolitan area (ascertained using Google maps; within x metres of a city centre), and the proximity of an English variety to a city centre (ascertained using google maps) will also be included. As will the area income (ascertained from web searches).

Frequency of papers will be measured using the search protocol outlined in the following subsection.

2.2 Search protocol for papers

Searches will be conducted in Google Scholar, and will be repeated in the databases of several linguistics journals concerned with documenting language variation and change. These databases will include the database of the journal Language Variation and Change, The Journal of Sciolinguistics and Linguistics Vanguard.

The search terms used will follow the formula, where 'name of variety' would be replaced with the wikipedia entry name for the variety of English, e.g. 'Geordie' and any alternative terms used for the same variety as suggested by wikipedia. The following searches will be conducted for each term found for each variety of English included in the study:

- 'name of variety' varia*
- 'name of variety' sociolinguist*

Only a single term for each variety of English will be used in searches so as not to bias the variable 'frequency of papers' in the process of literature searches.

Once all searches have been conducted, abstracts will be screened to assess whether they are emprical studies of variationist sociolinguistic phenomena. Frequency of papers from each search after all non-relevant studies have been removed will be included in the final analysis dataset.

how do you think we should handle variation in terminology - e.g. Geordie vs. Tyneside English (though notably, this term variation is actually mentioned on the wiki page)

2.3 Search protocol for corpora and the existence of research institutions

The following search terms will be used to assess the existence of corpora and research institutions:

- 'name of variety' corp*

- 'name of variety' lab
- 'name of variety' research

2.4 Inclusion criteria

Where databases don't allow wildcards, the first search term will be changed to *'name of variety' corpus*.After searches have been conducted, datasets will be downloaded and scanned to ensure that articles meet the following inclusion criteria:

- Studies must be empirical assessments of phonetic change in apparent time.
- Studies

Any studies that do not meet these criteria will be removed from datasets.

2.5 Measuring geography

Distances between geographical locations will be measured in R using ggmap.

2.6 Statistical analysis

Linear models will be used to test all variables as predictors of frequency of publications. These models will be constructed using R (R Core Team 2018).

The most current available datasets and statistical analysis can be found on the Open Science Framework (https://osf.io/bp3es).

References

Kahle, D. & Wickham, H. (2013), 'ggmap: Spatial visualization with ggplot2', *The R journal* **5**(1), 144–161.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Trudgill, P. & Watts, R. J. (2002), *Alternative histories of English*, Routledge.

Wells, J. C. (1982), *Accents of english*, Vol. 2, Cambridge University Press Cambridge.