Liam Keeble[1] and Caitlin Halfacre[2]

[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

Corresponding author

Assessing research bias against English varieties: a systematic review

Abstract

Keywords: Research bias, Bibliometrics, Sociolinguistics

2010 MSC: 00-01, 99-00

1. Introduction

This paper will ask several questions:

- Are English varieties that are typically geographically distant from linguistics university departments understudied?
- Does the presence of a locally focused corpus increase the research output on a particular variety?
- Are English varieties associated with higher social/income status lacking in research articles?
- Is most research conducted on varieties of English typically associated with suburban, as opposed to 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

Preprint submitted to Elsevier *March 1, 2021*

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 [1, 2]. 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

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 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:

how do you think we should handle variation in terminology - e.g. Geordie vs. Tyneside English (though

- 'name of variety' varia*

- ‘name of variety’ sociolinguist*

Once all searches have been conducted, abstracts will be screened to assess whether they are emprical studies of variationist sociolinguistic phenomena. Frequency of papers included in the final dataset for each variety of English will be included in the final analysis dataset.

2.3. Statistical analysis

Linear models will be used to test all variables as predictors of frequency of publications. These models will be constructed using R [2].

The most current available datasets and statistical analysis can be found on the Open Science Framework.

3. Results

4. Discussion

References

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

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