

# Assessing research bias against English varieties: a systematic review

Caitlin Halfacre<sup>2</sup> and Liam Keeble<sup>1</sup>

<sup>1</sup>Henry Wellcome Building, Medical School, Newcastle Upon Tyne,  
NE2 4HH, United Kingdom

<sup>2</sup>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.

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

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

## 55 2.2 Search protocol for papers

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

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

- 65 • 'name of variety' varia\*
- 66 • 'name of variety' sociolinguist\*

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

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

## 73 2.3 Search protocol for corpora and the existence of research 74 institutions

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

- 77 • 'name of variety' corp\*

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)

- 78       • ‘name of variety’ lab
- 79       • ‘name of variety’ research

## 80   **2.4   Inclusion criteria**

81   After searches have been conducted, datasets will be downloaded and scanned to  
82   ensure that articles meet the following inclusion criteria:

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

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

## 86   **2.5   Measuring geography**

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

## 88   **2.6   Statistical analysis**

89   Linear models will be used to test all variables as predictors of frequency of publi-  
90   cations. These models will be constructed using R (R Core Team 2018).

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

## 93   **References**

- 94   Kahle, D. & Wickham, H. (2013), ‘ggmap: Spatial visualization with ggplot2’, *The*  
95       *R journal* **5**(1), 144–161.
- 96   R Core Team (2018), *R: A Language and Environment for Statistical Computing*,  
97       R Foundation for Statistical Computing, Vienna, Austria.  
98       **URL:** <https://www.R-project.org/>
- 99   Trudgill, P. & Watts, R. J. (2002), *Alternative histories of English*, Routledge.
- 100   Wells, J. C. (1982), *Accents of english*, Vol. 2, Cambridge University Press Cam-  
101       bridge.