

Assessing research bias against English varieties: a systematic review

Caitlin Halfacre² and Liam Keeble¹

¹Henry Wellcome Building, Medical School, Newcastle Upon Tyne,
NE2 4HH, United Kingdom

²Percy 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 jiiiiii HEAD Searches will be conducted in Google Scholar, and will be repeated
57 in the databases of several linguistics journals concerned with documenting lan-
58 guage variation and change. These databases will include the databases for English
59 Language and Linguistics, Language Variation and Change, The Journal of Soci-
60 olinguistics and Linguistics Vanguard.

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

- 66 • Search Term: (WC=(Linguistics) AND ((ALL="name of variety") AND ((ALL=sociolinguist*)
67 OR (ALL=varia*) OR (ALL=change))))
- 68 • Document Type: All
- 69 • Timespan: 1982-2019

70 Only a single term for each variety of English will be used in searches so as
71 not to bias the variable 'frequency of papers' in the process of literature searches.
72 Searches will be conducted in the Web of Science database. The search terms used
73 will follow the formula, where 'name of variety' would be replaced with the wikipedia
74 entry name for the variety of English, e.g. 'Tyneside'. Thus, the following search
75 will be conducted for a single term found for each variety of English included in
76 the study: 'name of variety' English. Only a single term for each variety of English
77 will be used in searches so as not to bias the variable 'frequency of papers' in the
78 process of literature searches. The first 50 results will be downloaded from each
79 search.

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)

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

84 **2.3 Search protocol for corpora and the existence of research** 85 **institutions**

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

- 88 • 'name of variety' corp*
- 89 • 'name of variety' lab
- 90 • 'name of variety' research

91 **2.4 Inclusion criteria**

92 After searches have been conducted, datasets will be downloaded and scanned to
93 ensure that articles are studies of sound change in apparent time, with the following
94 inclusion criteria:

- 95 • the study is empirical
- 96 • the dependent variable is phonetic or phonological
- 97 • the study assesses change
- 98 • Studies must be studying the appropriate variety of English.

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

100 **2.5 Measuring geography**

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

102 **2.6 Statistical analysis**

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

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

107 **References**

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

110 R Core Team (2018), *R: A Language and Environment for Statistical Computing*,
111 R Foundation for Statistical Computing, Vienna, Austria.

112 **URL:** <https://www.R-project.org/>

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

114 Wells, J. C. (1982), *Accents of english*, Vol. 2, Cambridge University Press Cam-
115 bridge.