Integrating Aggregational and Probabilistic Approaches to Dialectology and Language Variation

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
der Philologischen Fakultät
der Albert-Ludwigs-Universität
Freiburg i. Br.

vorgelegt von

Christoph Benedikt Sebastian Wolk
aus Freiburg i. Br.

WS 2013/2014

Erstgutachter: Prof. Dr. Benedikt Szmrecsanyi
Zweitgutachter: Prof. Dr. Bernd Kortmann

Vorsitzender des Promotionsausschusses
der Gemeinsamen Kommission der
Philologischen, Philosophischen und Wirtschafts-
und Verhaltenswissenschaftlichen Fakultät: Prof. Dr. Bernd Kortmann

Datum der Disputation: 31.03.2014

Contents

List of Maps v

List of Tables vii

List of Figures ix

List of Abbreviations xi

List of County codes xiii

1. Introduction 1
- 1.1. Frequency and variation . . . 1
- 1.2. The dialect landscape of Britain . . . 7
- 1.3. Outline . . . 11

2. Aggregation and Language Variation 13
- 2.1. Aggregational methodologies: an introduction . . . 14
  - 2.1.1. Calculating linguistic distance . . . 16
  - 2.1.2. Fundamental distance analysis methods . . . 20
- 2.2. Applications of aggregational techniques . . . 30
  - 2.2.1. Dialectometry . . . 30
  - 2.2.2. Aggregational techniques in historical linguistics . . . 33
- 2.3. Corpus-based dialectometry . . . 36

3. Data and Methods 39
- 3.1. Data . . . 39
  - 3.1.1. The Freiburg Corpus of English Dialects . . . 39
  - 3.1.2. The feature set . . . 46
- 3.2. Methods . . . 46
  - 3.2.1. Corpus-based dialectometry . . . 46
  - 3.2.2. Extending corpus-based dialectometry . . . 49
  - 3.2.3. Automated bottom-up syntactic classification . . . 64
- 3.3. Chapter summary . . . 75

4. Feature-based analyses 77
- 4.1. Model-based analyses . . . 77
  - 4.1.1. Single feature models . . . 77
  - 4.1.2. On the effect of additional predictors . . . 137

4.1.3. Sociolinguistic summary . . . 148
4.1.4. Geolinguistic summary . . . 155
4.2. POS n-gram-based analyses . . . 166
4.2.1. Geolinguistic results: unigrams . . . 166
4.2.2. Geolinguistic results: bigrams . . . 168
4.2.3. Sociolinguistic results: gender . . . 180
4.2.4. Sociolinguistic results: age . . . 186
4.3. Chapter summary . . . 191

5. Aggregational analyses 193
5.1. Comparing normalization- and model-based approaches . . . 194
5.2. Dialect areas: hierarchical clustering . . . 197
5.2.1. Normalization-based results . . . 198
5.2.2. lmer-based results . . . 200
5.2.3. GAM-based results . . . 203
5.2.4. Bottom-up analysis: frequency . . . 204
5.2.5. Bottom up analysis: reliability . . . 205
5.2.6. Interim summary . . . 209
5.3. Intersecting dialect areas and continua . . . 210
5.4. Chapter summary . . . 217

6. Discussion & Conclusion 219
6.1. Summary . . . 219
6.2. What do we gain? . . . 221
6.2.1. On the influence of data availability . . . 222
6.2.2. On geographic and linguistic distances . . . 227
6.2.3. On non-geographic factors . . . 233
6.2.4. On bottom-up versus top-down analysis . . . 234
6.3. On the dialect landscape of Britain . . . 237
6.3.1. Revisiting the outliers . . . 237
6.3.2. The North of England . . . 240
6.3.3. The South and Midlands . . . 245
6.4. Outlook and concluding remarks . . . 249

Appendices 255

A. CLAWS7 Tag Set 255

B. Technical notes 261
B.1. Top-down . . . 261
B.1.1. Data . . . 261
B.1.2. lmer models . . . 261
B.1.3. GAMs . . . 262
B.1.4. Aggregation . . . 262

B.2. Bottom-up 263
B.2.1. Normalization 263
B.2.2. Counting and permutation 264
B.2.3. Analysis 265

Bibliography **269**

Deutsche Zusammenfassung **287**

List of Maps

1. Dialect area classifications . . . 9

2. FRED overview . . . 45
3. GAM map for Feature 5 . . . 66
4. Geographic distribution of the bigram `PPH1.VBDR`, *it were* . . . 74

5. Single feature models: Features 1–3 . . . 81
6. Single feature models: Features 4–5 . . . 82
7. Single feature models: Feature 6 . . . 85
8. Single feature models: Features 7–9 . . . 86
9. Single feature models: Features 11 & 12 . . . 89
10. Single feature models: Features 13–14 . . . 90
11. Single feature models: Features 15 & 16 . . . 93
12. Single feature models: Features 17–20 . . . 94
13. Single feature models: Features 21–23 . . . 99
14. Single feature models: Features 24 & 25 . . . 100
15. Single feature models: Feature 26 . . . 103
16. Single feature models: Features 27 & 28 . . . 104
17. Single feature models: Features 29 & 30 . . . 107
18. Single feature models: Features 31 & 32 . . . 108
19. Single feature models: Features 33–35 . . . 113
20. Single feature models: Features 36–38 . . . 114
21. Single feature models: Features 39–41 . . . 119
22. Single feature models: Features 42 & 43 . . . 120
23. Single feature models: Features 44 & 45 . . . 125
24. Single feature models: Features 46 & 47 . . . 126
25. Single feature models: Feature 48 . . . 129
26. Single feature models: Features 49 & 50 . . . 130
27. Single feature models: Features 51–54 . . . 133
28. Single feature models: Features 55–57 . . . 134
29. Case study: contraction in negative contexts . . . 145
30. Bigrams involving *was* and *were* . . . 171
31. Bigrams: *used to* . . . 172
32. Bigrams involving *them* . . . 177
33. Bigrams: -*nae* . . . 177
34. Bigrams involving *do* . . . 178

35. Comparison of methods: normalization, lmer and GAM modeling 196
36. Cluster analysis based on normalized feature frequencies 201
37. Cluster analysis based on normalized feature frequencies, FRED-S only . . 201
38. Cluster analysis based on lmer model predictions 202
39. Cluster analysis based on GAM predictions 202
40. Cluster analysis based on unigram frequencies 207
41. Cluster analysis based on bigram frequencies 207
42. Cluster analysis based on unigram reliability scores 208
43. Cluster analysis based on bigram reliability scores 208
44. Continuum maps 216

List of Tables

2.1. Example data illustrating the Hamming distance measure 18
2.2. Example data set: data matrix 22
2.3. Example data set: distance matrix 22
2.4. Example data set: distance matrix after first clustering iteration 23
2.5. Example data set: distance matrix after second clustering iteration 23
2.6. Data for the network example 27

3.1. Feature set used for the model-based analyses 47
3.2. Distance matrix calculation 49
3.3. lmer model for Feature 5: *us* 57
3.4. Comparison of normalization and lmer modeling: simulation results 61
3.5. Bottom-up statistics for the bigram `PPH1.VBDR`, *it were* 73

4.1. Contraction: lmer model results 142
4.2. Contraction: GAM results 143
4.3. Contraction: model predictions compared to reported values 147
4.4. Summary of the effects of gender across models 149
4.5. Summary of the effects of speaker age across models 152
4.6. Summary of the interaction of speaker age and gender across models . . . 154
4.7. Summary of geographical distribution characteristics 156
4.8. Largest lmer county effect variances 159
4.9. Largest and smallest significant GAM effective degrees of freedom 159
4.10. Highest significant Moran's *I* values for lmer county predictions 160
4.11. Highest significant Moran's *I* values for GAM predictions 161
4.12. Associations of feature bundles to FRED regionalizations 162
4.13. Most relevant unigrams 167
4.14. Most relevant *was/were* related bigrams 169
4.15. Most relevant related bigrams related to *used to* 173
4.16. Most relevant bigrams related to *them* 174
4.17. Most relevant bigrams related to -*nae* 176
4.18. Most relevant bigrams related to -*do* 180
4.19. Unigrams by gender: nouns and pronouns 182
4.20. Bigrams by gender: nouns and pronouns 183
4.21. Unigrams by gender: verbs 184
4.22. Bigrams by gender: verbs 185
4.23. Unigrams by gender: other 186
4.24. Bigrams by gender: other 187

4.25. Unigrams by age: nouns and pronouns 189
4.26. Bigrams by age: nouns and pronouns 189
4.27. Unigrams by age: verbs 190
4.28. Bigrams by age: verbs 190
4.29. Unigrams by age: other 191
4.30. Bigrams by age: verbs 191

6.1. Features associating Northumberland with the Scottish Lowlands or the North of England 242
6.2. Feature bundles and the North 242
6.3. Features associating Lancashire with the North of England or the Midlands 243
6.4. Distinctively Southwestern Bigrams 246

List of Figures

2.1. Two-dimensional MDS analysis of example data set 21
2.2. Clustering result: single linkage 24
2.3. Clustering solutions using other linking algorithms 25
2.4. Phenograms for the example data set 26
2.5. Splits graph example 28

3.1. Distribution of normalized frequencies: *us* 50
3.2. BLUPs (intercept adjustments) by county for Feature 5: *us* 57
3.3. Visualization of simulation results 62
3.4. GAM perspective plot for *us* 65

4.1. Aggregate view on the feature distribution: Hierarchical cluster plot 165

5.1. Comparison of methods: normalization vs. lmer 195
5.2. Splits graph based on distances derived from lmer model predictions . . . 211
5.3. Splits graph based on distances derived from GAM predictions 212
5.4. Splits graph based on bigram reliability distances 215

6.1. Linguistic distance and corpus size differences 223
6.2. Linguistic distance and minimal corpus size 226
6.3. Linguistic distance and geographic distance 232

List of Abbreviations

BLUP	best linear unbiased predictor
CBDM	corpus-based dialectometry
DM	dialectometry
edf	estimated degrees of freedom
FRED	Freiburg Corpus of English Dialects
FRED-S	Freiburg Corpus of English Dialects Sampler
GAM	generalized additive model
GCV	generalized cross validation
GLMM	generalized linear mixed model
lmer	function name in the statistical package *lme4* for R; used here as a shorthand for GLMMs created with lmer
LOESS	locally weighted scatter-plot smoothing
MDS	multi-dimensional scaling
NJ	Neighbor-Joining
NORM	non-mobile older rural man
PERL	a scripting language, sometimes glossed as Practical Extraction and Report Language
POS	part-of-speech
pttw	per ten thousand words
RGB	red, green, blue color space
SED	Survey of English Dialects
UPGMA	unweighted pair group method with averaging
VARBRUL	variable rules analysis
WAVE	World Atlas of Variation in English (Kortmann & Lunkenheimer 2013)
WPGMA	weighted pair group method with averaging

List of County codes

ANS	Angus
BAN	Banffshire
CON	Cornwall
DEN	Denbighshire
DEV	Devon
DFS	Dumfriesshire
DUR	Durham
ELN	East Lothian
GLA	Glamorganshire
HEB	Hebrides
KCD	Kincardineshire
KEN	Kent
LAN	Lancashire
LEI	Leicestershire
LND	London
MAN	Isle of Man
MDX	Middlesex
MLN	Midlothian
NBL	Northumberland
NTT	Nottinghamshire
OXF	Oxfordshire
PEE	Peebleshire
PER	Perthshire
ROC	Ross and Cromarty
SAL	Shropshire
SEL	Selkirkshire
SFK	Suffolk
SOM	Somerset
SUT	Sutherland
WAR	Warwickshire
WES	Westmorland
WIL	Wiltshire
WLN	West Lothian
YKS	Yorkshire

Acknowledgments

Many have supported me in my work, and I am deeply indebted to them. In particular, I would like to thank:

- Benedikt Szmrecsanyi, my supervisor, for his support, feedback, coaching and kindness. Without his help, none of this would have been possible.

- My collaborators on various projects, in particular Bernd Kortmann, Joan Bresnan, Anette Rosenbach, Lars Konieczny, Katharina Ehret, Peter Baumann, Sascha Wolfer, Daniel Müller-Feldmeth and Tobias Streck for many interesting discussions and for broadening my mind.

- The Freiburg Institute for Advanced Studies (FRIAS) for its generous support and a stimulating environment.

- The *Graduiertenkolleg DFG GRK 1624 "Frequenzeffekte in der Sprache"*, where I had the honor of being an associated member.

- The audiences at the GRK 1624 colloquia, the Freiburg Workshop *Quantitative Methods in Geolinguistics*, the *Methods in Dialectology XV* in Groningen and the *ISLE3* in Zurich, and in particular John Nerbonne and Joan Bresnan, for their helpful comments and questions.

- The authors of the software used in this analysis: to Peter Kleiweg for *RuG/L04*, Daniel Huson and collaborators for SplitsTree, and the many contributors to R.

- Katharina Ehret for help with proof-reading.

- My friends and family, especially my parents Ingrid and Arnold Wolk, my siblings Simon, Helena and Clarissa, my grandparents Ottilie and Franz, my aunt Brigitte and my uncle Rolf.

All remaining errors and omissions are my own.

1. Introduction

1.1. Frequency and variation

The present work aims to advance the methodological tool set for the analysis of dialectal variation; more specifically, it concerns itself with analyses based on naturalistic corpus data. It is a direct successor of pioneering work by Szmrecsanyi (2008; 2013) in a framework named *corpus-based dialectometry* (CBDM). The central characteristics of this approach – as I employ it here – can be summarized by the following characteristics:

- centered on morphosyntax
- corpus-based
- frequency-driven
- aggregational

I will discuss these in turn.

First, CBDM is *centered on morphosyntax.* Throughout its history, dialectology has tended to focus on lexis and pronunciation. Although Kirk (1985: 130) notes a consensus among several Edinburgh dialectologists "that it is from grammatical material, especially the syntactical, that the most interesting results for linguistic variation are to be expected", most large-scale collections of dialectal data, and the atlases resulting from them, have detailed coverage for lexical and phonetic differences, but are relatively sparse for morphological and syntactic features. There are two major reasons for this (Ihalainen 1988: 569). One is theoretical: Szmrecsanyi (2013: 159) notes that some scholars consider morphosyntax less *raumbildend* than lexicon or phonology and its geographic variation less salient; as examples of this he references, among others, Lass (2004: 374)[1] and Wolfram & Schilling-Estes (1998: 161). Ihalainen (1988) quotes (Wakelin 1977: 125) as a proponent of the similar idea that there is "in general an underlying identity of syntactical patternings in all forms of English".

[1] Lass does, however, note that there are exceptions to this, such as Scots and the Southwest of England.

Other authors note that grammar is not as easily studied using the survey-based method of traditional dialectology. Consider the three possible realizations of the dative construction in British English dialect grammars - recipient first, recipient second with *to*, and recipient second without *to*, as in (1) (see also Section 4.1.1.11.2).

(1) a. I gave him it.
b. I gave it to him.
c. I gave it him.

Comparing this to lexical variation, Kirk (1985: 133) notes:

> Syntax is different. Most speakers could readily produce all three of the mapped variants [...] whereas no speaker (apart from schooled dialectologists) would be likely to share *dolly posh* and *draidlock* in their vernacular vocabularies.

This poses obvious challenges to data collection – simply relying on informants' individual judgment is likely to lead to heavily distorted results. Ihalainen (1988) argues that data of both a different kind (namely, tape-recorded speech) and different volume (large quantities) will be necessary to properly deal with dialect syntax; finally, new methods for dealing with such data will need to be developed. Nevertheless, many traditional surveys include at least some morphosyntactic features, and some specialized investigations into morphosyntax using such methodology were carried out, such as the *Survey of British Dialect Grammar* (Cheshire 1989).

Recent years have "[witnessed] on a broad scale an increasing interest in dialect grammar" (Kortmann 2004b: 3). This interest stems from several angles, and takes different forms. One important aspect is the growing body of research into post-colonial Englishes (Schneider 2007) and their developmental history (Hickey 2004). Especially for morphosyntactic phenomena, features of British English dialects turned out to be quite understudied, making it difficult to trace to what degree a given feature distribution in a "new English" is an influence from a British founding dialect or an independent development. The question of "[h]ow [...] the roots of communities and regions and countries play out in the way their dialects are used by contemporary speakers several hundred years later" (Tagliamonte 2013: xii) led to a large number of studies by Sali Tagliamonte's research group (e.g. Tagliamonte & Lawrence 2000, Tagliamonte & Smith 2002, Tagliamonte et al. 2005, Tagliamonte & Baayen 2012). While these studies generally focus on a small number of locations, they utilize the full methodological apparatus of modern variationist sociolinguistics to provide detailed accounts of the features under study. Around the same time, typological databases of non-standard varieties of English were compiled, first as part of the *Handbook of Varieties of English* (Kortmann et al. 2004) including both phonology and morphosyntax, and later and specifically for morphosyntax as the *World Atlas of Variation in English* (Kortmann & Lunkenheimer 2013). This allowed the investigation of the typological distribution of features in Englishes world-wide using a quantitative approach (e.g. Szmrecsanyi & Kortmann 2009, Kortmann & Wolk 2013). These studies investigate how different morphosyntactic features pattern across the world and across types of Englishes. Even formal approaches to grammar, a domain which previously tended to ignore non-standard structures, have begun to explore "how theoretical modelling can be enriched by taking variation as a core explanandum" (Adger & Trousdale 2007: 274). This general increase of interest has coincided with a crucial development: the kind of data that Ihalainen (1988) envisioned has become available, and this leads us to the next characteristic of the approach presented here.

CBDM is *corpus-based.* As noted above, the primary data sources for most dialectological work are dialect atlases, typically compiled by fieldworkers using a questionnaire-based method. Szmrecsanyi (2013: 4) therefore considers the atlas signal to be "analytically twice removed (via fieldworkers and atlas compilers) from the analyst". A dialect corpus, i.e. a large collection of natural dialect speech, is a more direct source of linguistic information and has several beneficial properties: first, the research questions are not constrained by the questionnaire design. As long as the feature of interest is frequent enough for the amount of linguistic material available, it can be analyzed by the corpus user, even if those collecting the data did not explicitly choose to support that particular feature. Second, as Szmrecsanyi (2013: 4) notes, the data elicited by questionnaires is often "meta-linguistic" in nature, as a response to a fieldworker's question concerning the informant's language use. It is not guaranteed that this matches the informants' linguistic behavior in more natural settings. Corpora, on the other hand, are records of exactly such behavior. Finally, the corpus signal allows a different type of information, which leads us to the third major characteristic of the CBDM approach.

Corpora allow a *frequency-driven* approach to linguistic analysis. The atlas signal is in essence categorical, answering, for a given location, questions of the type: *How is a given word pronounced? What words do speakers use? Which grammatical constructions are allowed?* These questions can only represent part of the linguistic reality, as they necessarily hide the gradient properties of variation that may exist. Using a corpus, the analyst can not only find out what is available, but can, given enough data, precisely determine how often it is used and under which circumstances. This can be relevant for all linguistic levels, but is especially so for morphosyntax. As we have seen from Kirk's quote above, grammatical features are especially difficult to handle using questionnaires. Frequency information can solve this problem: most speakers can and will produce different realizations of a grammatical phenomenon on individual occasions, but the sum of many observations yields a more accurate picture of how speakers of a given dialect behave.

Compared to the written material that fills the bulk of modern mega-corpora such as the *British National Corpus* or Davies's *Corpus of Contemporary American English* (2008-), spoken corpora are much more labor-intensive to compile. The restriction to dialectal material adds further layers of difficulty. Nevertheless, dialect corpora of respectable sizes have become available. In English, this includes, among others, the *The Helsinki Corpus of British English Dialects* (2006) of about 1 million words and the 2.5 million words strong *Freiburg Corpus of English Dialects* (FRED). To add an example for non-English dialect corpora: the *Nordic Dialect Corpus* (Johannessen et al. 2009) is a collection of subcorpora containing dialect material from six North Germanic languages, spanning about 2.8 million words. The availability of such corpora has led to a growing number of studies doing dialectology with a corpus methodology, including many conducted at the University of Freiburg on the basis of FRED (e.g. Wagner 2002, Herrmann 2003, Anderwald 2003, Schulz 2012).

The final characteristic of CBDM is that it is an *aggregational* approach. Traditional dialectology tends to focus on individual features and attempts to abstract their distribution into geographically meaningful patterns. The problem is that individual features often do not agree with one another; as in Bloomfield's famous dictum, "every word has its own history" (1933: 328). Single-feature analyses fall short when the object of interest is not a single characteristic, but the dialect 'as a whole'. The dialectometric approach attempts to solve this problem by considering a large number of features holistically. Even if each feature has its own history and distribution, taken together, they constitute the dialect as a whole. Investigating dozens – or thousands – of characteristics simultaneously thus leads to a more accurate description of a dialect in its relation to others. Dialectometry as a research project has a considerable tool set of analysis techniques and visualization types. A more detailed description and explanation follows below as Chapter 2. Until the end of the last decade, the frequency-based investigation of regional variation in morphosyntax and the dialectometric approach to feature aggregation were separate projects, and dialectometric analyses generally tapped classic dialect atlases as their data source. Since then, various approaches have made considerable progress in marrying the "jeweler's-eye perspective" of quantitative corpus analysis with the "bird's-eye perspective" of dialectometry (Szmrecsanyi 2013: 2), of which CBDM is one. Other notable investigations are those by Grieve (2009) and Sanders (2010), which will be discussed in

Section 2.3.

In this dissertation, I attempt to move this union forward. My approach has three major characteristics in addition to the four discussed so far. It is:

- probabilistic
- both top-down and bottom-up
- incorporating sociological information

First, my approach is *probabilistic.* The major impetus lies in the following: I fully agree with Szmrecsanyi (2013: 163) that "frequency noise [is] part of linguistic reality". Frequency noise is, however, also part and parcel of frequency-based investigations themselves. It is, in general, not easy to determine whether the observed noise represents true variability in the signal or is actual noise, i.e. an artifact of the individual data set and its composition. This is especially troubling when the uncertainty is high or unevenly distributed - a measurement that is based on a small sample will generally be less accurate than one based on a larger sample, and comparing the two as if they were the same will lead to wrong estimates. I will show, through conceptual arguments, simulations, and finally through a reanalysis of Szmrecsanyi's results, that his method fails to adequately take this into account. I will also explore ways in which this can be remedied. One involves building probabilistic models of the feature distributions in such a way that the influence of some biases can be removed – at the cost of introducing new ones.

Second, I complement the *top-down* approach with a *bottom-up* investigation. CBDM is a top-down approach that first defines the features under study, then bases its analysis on their frequencies. The bottom-up approach works in the other direction: it starts from the corpus in its part-of-speech tagged form and counts the syntagmatic sequences that appear. In this way, dialectologically interesting features are not presupposed, but emerge from the data. This also leads to much finer-grained features: Szmrecsanyi, for example, includes the primary verb *to do* in his feature list (Feature 13, see Section 4.1.1.3.1). The bottom-up approach includes separate counts for all forms of *to do* with their local context, such that the past participle *done* preceded by the nominative first person singular personal pronoun *I* is measured independently from, say, *do* preceded by a proper name. This approach has two goals: first, more fine-grained features should allow finer patterns to emerge. It also makes more of the available data usable, as each single word can enter the analysis, not only those preselected to be relevant. Thus, the bottom-up approach may also help to alleviate the problem of data availability.

Finally, the (top-down) analyses presented here can make use of additional information, which for present purposes pertains mostly to *sociological information*: speaker age and gender. Szmrecsanyi's method effectively treats all speakers as the same. Yet it is a consistent result in sociolinguistic analysis that female speakers tend to use fewer non-standard forms than male speakers do (Chambers 2003: 116). If one subcorpus contains more material by female speakers than another, it is not clear whether the resulting frequency differences stem from dialectal or gender differences. Probabilistic modeling can estimate the effect that gender and age have on the data as a whole, and therefore reduce such imbalances. The scope of additional information is also not limited to age and gender. I will present a more elaborate analysis of one feature as a case study, where several language-internal factors are taken into account.

The following research questions summarize the project:

- To what degree does the amount of available data influence the result of the measurement? Can the influence of this factor be reduced?
- If we can improve the measurement, how does this influence Szmrecsanyi's results concerning, for example, the relation between linguistic distance and geographic distance?
- Do non-geographic factors such as speaker age and gender play a role in the aggregational approach?
- How do "top-down" approaches compare with "bottom-up" approaches?
- What do the results from these methods tell us about the structure of morphosyntactic variation in the British Isles?

The top-down part reuses the data from Szmrecsanyi (2013), with additional checking and cleanup. Why do I reanalyze this instead of creating a new data set? There are three major reasons. Most importantly, it is of exceptional quality. It uses the FRED corpus, which is the largest available dialect corpus for British English, and Szmrecsanyi's feature catalog is comprehensive. Most other features are so rare that a quantitative analysis is not feasible even on a large corpus like FRED (cf. the list of excluded features in Szmrecsanyi 2013: 37). Furthermore, one of the explicit goals of the present work is to test the CBDM approach, and this is facilitated by a direct comparison with the flagship study in this paradigm. Finally, the original data set is publicly available[2]. This allows further methodological refinement, as future progress in CBDM can be directly compared against both Szmrecsanyi's results and the ones presented here.

The next section will provide some background on the existing reports of the geographical structure of British English dialect variability.

[2] It can be downloaded from `https://sites.google.com/site/bszmrecsanyi/datasets`

1.2. The dialect landscape of Britain

Many dialectologists have provided classifications of British English dialects into large-scale areas; a detailed discussion is given by Szmrecsanyi (2013: Chapters 1 and 6). In this section, I will provide a concise synthesis of the classifications found there as well as Szmrecsanyi's results, and add two more recent studies. All schemes detailed here and in Szmrecsanyi (2013) differ by their areal coverage: many do not include Wales or Scotland, or only include parts of these areas. As far as possible, the regions were matched to the counties included in this work (see Section 3.1.1). Some of the classifications are visualized in Map 1.

Baugh & Cable (1993) adopt a historical perspective and provide two classifications, one each for the dialect areas of Old and Middle English (Maps 1a and 1b). The Old English scheme contains four groups and that for Middle English contains five; both cover England and part of the Scottish Lowlands. In general, both classifications are quite similar for the areas relevant to this study, although there are points of disagreement between the two. Both distinguish the West Saxon South(west) of England from the Kentish dialects, which span Kent, London, and, in OE, also Middlesex. The Mercian Midlands of OE, covering the Midlands, East Anglia, and Oxfordshire are divided into western and eastern parts in the ME classification, with Middlesex falling into the eastern group. The dialects north of the river Humber, including the Scottish dialects that are covered in their analysis, form the Northumbrian group in OE. In the ME scheme, this group is labeled Northern and excludes Lancashire, which here belongs to the West Midlands.

Ellis (1889) provides a classification based on his extensive survey of English dialects, resulting in 42 areas and 6 major groups, 5 of which appear in the data analyzed here (Map 1c). Ellis places the Southeast and the Southwest of England together as the Southern group, while East Anglia and Middlesex form the Western group. The Midlands, the North of England, and the Scottish Lowlands constitute groups of their own. There are many classifications that draw upon the monumental *Survey of English Dialects* (SED), conducted by Orton & Dieth (1962), and the various interpretations that were published as atlases, such as the *Linguistic Atlas of England* (Orton et al. 1978) or the *Structural Atlas of the English Dialects* (SAED, Anderson 1987).

A particularly influential one is Trudgill (1999), who provides both a 'traditional' and a 'modern' classification. Both are based on pronunciation differences, using a careful selection of dialectologically relevant features. Similar to Ellis, he finds six groups, which form two major areas: the South, including the Southeast and the Southwest of England as well as the Eastern and Western Midlands, and a northern group containing the North of England and the Scottish Lowlands. One point of contention between the traditional and modern schemes concerns Lancashire. As in Baugh & Cable's ME classification, Trudgill's traditional scheme includes Lancashire not as part of the North, but of the Western Midlands. In contrast, in the analysis of modern dialects (Map 1d), Lancashire is grouped with the North.

Goebl (2007a), drawing on the SED for a dialectometric analysis and covering only England, arrives at 8 distinct groups (Map 1e). His scheme makes a distinction between the Southeast[3] and the Southwest of England. Shropshire, part of the Midlands in many other classifications, here lies in the Northern Southwest. The Midlands themselves are divided into three groups: the Western Central, Eastern Central and Central East dialects. Finally, in the North, the dialects of Northumberland form their own group separate from the other Northern dialects. Again, Lancashire is not part of the North.

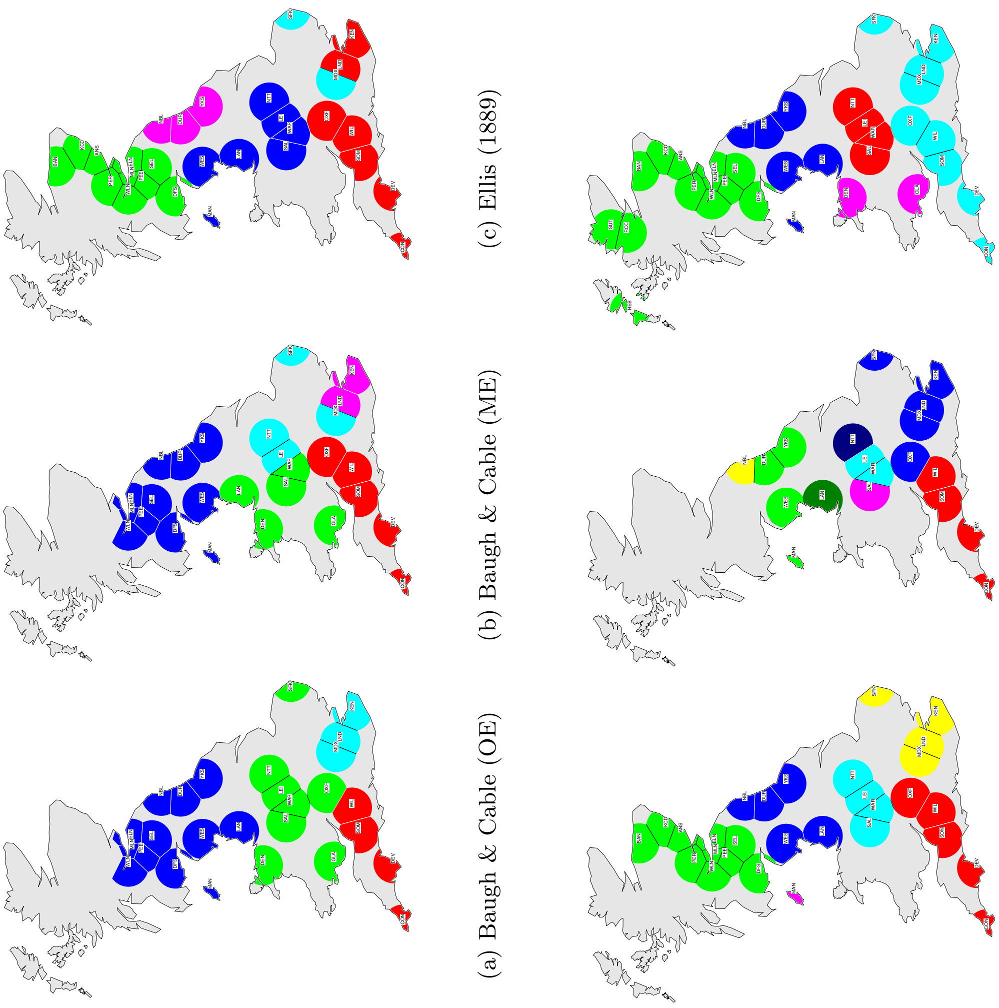
Inoue (1996) derives five dialectal areas by means of an experimental study in perceptual dialectology (Map 1f). His study includes Wales and Scotland, and both emerge as separate groups in his classification. England is divided into Southern, Northern and Midlands dialects; the North includes Lancashire.

Shackleton (2007; 2010) provides an analysis based on phonetic realizations derived either directly from the SED material as feature structures of a small selection of individual words, or from the classifications into phonetic variants in the SAED. He finds that the most important split separates the South from the Midlands, followed by a separation of the South into eastern and western parts, a separation of the Midlands from the North, and a segmentation of the North into three areas. Overall, his results are similar to those of Trudgill; the precise locations of the boundaries differ somewhat.

Szmrecsanyi (2013) provides two different schemes based on a dialectometric analysis using morphosyntactic data derived from the FRED corpus. Both result in geographically slightly different and discontinuous groupings. There is a tendency, however, toward having three large-scale groups: the South of England (plus Durham and Nottinghamshire as outliers), a group containing most varieties in the North of England and some of the Midlands, and finally a Scottish group that also contains Northumberland. The Midlands do not appear as a group separate from the North, which also includes Lancashire. Szmrecsanyi (127) reports that his results are statistically closest to the categorization by Ellis (1889).

Kortmann (2013), analyzing the *World Atlas of Variation in English* data, provides a classification of the areas of the British Isles based on feature frequency judgments by

[3] In analogy to Szmrecsanyi (2013: 9), I give the classifications in Goebl (2007a) names according to the scheme by Trudgill (1999).



(a) Baugh & Cable (OE)

(b) Baugh & Cable (ME)

(c) Ellis (1889)

(d) Trudgill (1999, Modern)

(e) Goebl (2007)

(f) Inoue (1996)

Map 1: Overview of several different dialect area classifications.

experts. His network-based representation emerges with four major zones: the Southeast and East Anglia are one, Ireland and the Isle of Man another, and Scotland the third. The final group comprises the Southwest, Wales, and the North of England. The position of the North here is quite curious and hard to explain. Nevertheless, Kortmann shows that for many individual features, broadly Northern and broadly Southern varieties exhibit clearly different patterns, and that the number of features that are characteristic of the North is higher than for the South.

In a recent study based on the *BBC Voices* project, a large-scale investigation into current linguistic variation across all of Great Britain, Wieling et al. (2013) focus on the lexical information gathered from the interactive web site of the project. In contrast to traditional dialectological work, their informants were therefore overall quite young. They consider the ten most frequent lexical variants for each of 38 concepts, and derive dialect areas based on the variant frequencies per British post code using bipartite spectral graph partitioning. They find that the major split runs between Scotland on the one hand, and England, Wales and Northern Ireland on the other. The next split involves the separation of a rather small partition of the far Scottish Northeast from the main Scottish group. For the non-Scottish dialects, the next division separates an area that corresponds to the North of England from the other dialects, with the border running south of Lancashire and Yorkshire.

As a summary, all dialect classifications overlap to a considerable degree. Crucially, all of them clearly distinguish the North of England from the South. Furthermore, several of the lower-level divisions, such as the one between the Southeast and the Southwest of England, are in principle similar between many schemes. Nevertheless, there are notable points of disagreement:

- Does Northumberland belong to the North of England, should it be grouped with Scotland, or is it a group of its own?
- Does Lancashire group with the North of England or with the Midlands?
- Do the Midlands constitute zero, one or multiple groups?
- How relevant is the distinction between the Southeast and the Southwest of England?

Section 6.3 will revisit these questions in the light of the new data and methods proposed in the present work.

1.3. Outline

Chapter 2 will introduce the basic ideas behind the aggregational approach to language variation. This will include some basic methodological concerns, such as the question of how the analyst is to proceed in establishing linguistic distances from different types of data. The aggregational analyses used in this work will also be introduced and explained here. This section will then continue with a discussion of dialectometry, the field that applies this view to dialectological data. Similar approaches in historical linguistics will also be introduced. A discussion of three approaches that apply dialectometric methods to corpus-based data will conclude the chapter.

Chapter 3 will first introduce the data set used in the present work: the dialect corpus FRED and the part-of-speech tagged version of its subcorpus FRED-S. Next, the original methodology by Szmrecsanyi will be presented. A discussion of potential problems with this approach will follow, concentrating on the influences of low data availability at some locations and on the possibility that factors external to geography may have an influence on the corpus-based frequencies. Two methods will be proposed that can, to some degree, address these concerns: *mixed-effects modeling* and *generalized additive modeling*. Next, innovative new methodology from Nerbonne & Wiersma (2006) and Sanders (2010) will be extended to the corpus at hand, measuring and evaluating syntactic distances on the basis of syntagmatic relationships between word classes.

Chapter 4 will apply these methods to a modified version of Szmrecsanyi's feature set and to the part-of-speech tagged version of FRED-S. For the two model-based approaches, each feature will be discussed individually, covering the feature itself, the extraction strategy, and the major results emerging from the models. This is followed by a case study of a single feature, the alternation between negative and auxiliary contraction. A more complex model will be used to explore how adding information concerning the local context of each token influences the results of the modeling process. I will then provide a synopsis of the influence of sociolinguistic factors on each feature, followed by a summary discussion of the patterns in geographical distributions. The focus will then switch to the *bottom-up* analyses and discuss n-grams that were uncovered as reliably different between counties. A presentation of the effect of social factors on the part-of-speech patterns will conclude this chapter.

Chapter 5 will leave the analysis of individual features behind and consider the data as a whole from the bird's eye perspective. First, the two model types introduced in Chapter 3 will be pitted against the normalization-based strategy used by Szmrecsanyi (2013) and against each other, to clarify the effect of each modeling strategy on the geolinguistic signal. Next, the distances resulting from the models as well as several bottom-up measures will be analyzed using hierarchical cluster analysis, to determine what areal signals can be found in the data. Finally, it will be investigated to which degree the data is consistent with the assumptions of a hierarchical areal structure. First, a splits graph representation using the NeighborNet algorithm will be used, then the structure of British English dialects as a continuum will be explored using a suitable cartographic representation.

The final chapter will begin with a summary of the work presented until then. I will then return to the research questions outlined in Section 1.1. It will be demonstrated that data availability is a significant influence on the results of corpus-based dialectometry, and that both modeling strategies can improve this somewhat, at the cost of introducing additional assumptions. I will also show that this influences Szmrecsanyi's results about the relationships between geographic distances, linguistic distances, and linguistic gravity. Finally, the role of sociolinguistic factors as well as the bottom-up oriented approaches in the corpus-based dialectometric enterprise will be discussed. Then, I will turn my attention back to the particular application of morphosyntactic variation in British English dialects. I will show that, despite the aforementioned problems, the core of Szmrecsanyi's analysis is confirmed, and that additional perspectives can highlight individual aspects of the multidimensional forest. I will conclude the work with a brief summary of the major results and a discussion of avenues for further development and research.

2. Aggregation and Language Variation

This section introduces the aggregational approach to linguistic variation in greater detail. Many sub-fields of linguistics deal with complex linguistic objects that may be characterized along multiple dimensions. Two of them are especially relevant for present purposes:

- dialectology concerns itself with dialects, spatially distributed varieties of the same language which vary from one another in a large number of features

- historical linguistics includes the grouping of languages into families based on their lexical, phonological and morphosyntactic similarities and differences

There are two major ways of dealing with such multi-dimensional objects: One approaches the object of study along a single dimension, i.e. an individual feature, in great detail, with the goal of achieving deep insights into the characteristics – whether distributional or developmental – of that specific feature. The other considers a large number of features with generally lower levels of detail, then utilizes a synoptic view of all these features to arrive at a holistic representation that makes the large-scale patterns of relations between the varieties or features under study more accessible.

Both of the approaches described above have useful applications, and each has a lot to offer to the other. I will discuss an advanced methodology for single-feature analysis, statistical models such as logistic regression, in the next chapter. The principles behind the aggregational view are of central importance to this work, and I therefore devote this chapter to how different sub-fields of linguistics utilize aggregational approaches for finding and visualizing patterns, with a focus on dialectometry. As there is considerable overlap, an introduction to aggregation will be given first, together with some general considerations. It will cover both the quantification of linguistic data and the most important statistical analyses used here. Then, aggregation in dialectology, i.e. dialectometry, will be discussed. While the present work explicitly does not attempt to do historical analysis, I will present one visualization technique commonly used in quantitative historical classification, splits graph representations, as has proven useful in several fields. Finally, three recent approaches to morphosyntactic dialectometry will be discussed.

2.1. Aggregational methodologies: an introduction

Let me begin by describing the key features of aggregational techniques as they are used in the present work. These methods are used to investigate the relationships between the *taxa*, the multidimensional objects under study, such as dialects, languages, or texts. The primary goal is not a strict test of previously formulated hypotheses, but rather exploratory data analysis – summarizing and visualizing the data to identify hidden patterns, leading to insights and the formulation of new hypotheses. Each object under study is evaluated with regard to multiple features, which together are considered to be representative of the total variability in the data as a whole. Usually, the number of feature dimensions is rather large, and can range from tens to thousands or more. Then, the individual measurements along these dimensions are viewed holistically, using a process that is strictly algorithmic and thus suitable for being automatized. While the process is fixed, this does not mean that the flow of the analysis strictly and mindlessly goes from raw data to the end result, nor that the expert knowledge of the researcher performing the analysis is dispensable. It does mean, however, that any adjustment has to be made explicit, either by formalizing it into the algorithm, by transformation of the primary data, or through the interpretation of the result.

Of course, a vast number of different methods exist that fit this description. New approaches that are especially adequate for a given application are continuously being developed, both within the linguistic domain and within other disciplines that employ exploratory data analysis – political science, evolutionary biology or community ecology to name just a few. Thus, a complete description is likely impossible, and certainly beyond the scope of the present work. The focus here will lie on presenting the characteristics and core issues that many of them share. The main purpose of this is to introduce the fundamental concepts required for the integration of aggregational and probabilistic analyses that will be constructed, tested, and applied in the later sections of this work. Background information, such as applications to categorical data, will offer important context and will be helpful in constructing extensions for different applications, even if the particulars are not immediately useful for the present purposes,

As defined above, aggregational methods begin with a set of taxa, whose number is customarily referred to as N, that are measured along a number of feature dimensions, p. This results in a data matrix $N \times p$, which contains the positions of each taxon in the feature space. The measurements in this matrix can be of different data types, three of which are particularly relevant for linguistic analysis. The first and simplest of these is the categorical data type, in which each measurement involves a judgment from a set of categories, i.e. a nominal scale. These categories may differ between feature dimensions and may range from a simple binary scheme, such as the presence or absence of a given feature in a language, to arbitrarily large sets, such as the set of lexical and phonetic realizations of a certain meaning in the varieties under study. The second data type is the string data type, linearly ordered sequences of characters using a fixed alphabet, such as Latin script or the International Phonetic alphabet. The advantage of this data type is that it is a natural representation for data concerning lexical or phonetic information, and that this allows for a more gradient representation of similarity between measurements – at the cost, however, of a more elaborate algorithmic process in analyzing the data. Finally, the third data type, and the one most relevant to this work, is the numeric data type, the natural representation for frequency data. These three types do not cover all aspects of linguistic data, and relevant methodologies for other applications continue to be developed. An example for this are measures for the comparison of syntactic trees (Sanders 2007, Noetzel & Selkow 1999 [1983]).

String and numeric data types can be reduced to the categorical data type, but this usually results in a loss of information. Trivially, strings can be used as categories, but this removes the similarity information between realizations that can be calculated from the string. For example, such an analysis is able to tell us that *eft* and *eff* are different from one another, but could not tell us that they are more similar to one another than they are to, say *padgetty poll* (see also Chambers & Trudgill 1998: 25ff.). A qualitative approach could identify categories based on etymology or, where the forms derive from the same ancestor, on individual character positions in the string that are considered interesting, such as whether the *t* of *eft* is present, and use each of these as a category in itself. Similarly, frequency could be made discrete by defining frequency categories and converting each number to the corresponding class. The benefit of this is that many methods of the aggregational framework can straightforwardly be applied to categorical data. Furthermore, the additional qualitative step allows the researcher to use prior knowledge to put emphasis on relevant features and to reduce noise. These advantages come at the risk of reduced accuracy.

We can distinguish two main types of aggregational analysis. Both start from the $N \times p$ matrix. One set of methods then explicitly aggregates over the features, resulting in a $N \times N$ (dis)similarity that can serve as the input for further analysis. This approach will be discussed below. The other set of methods eschews this step and begins the statistical investigation directly from the original input matrix. Character-based analyses of phylogeny inference, which will be discussed in Section 2.2.2, belong to this group. Another group of methods with this characteristic are those related to principle component analysis, such as factor analysis or the correspondence analysis family (e.g. Abdi & Valentin 2007). In these methods, the feature space is rotated with the goal of finding the configuration in which a small number of dimensions is maximally informative. In other words, if the data of two or more dimensions are strongly associated with one other dimension, the feature space is likely to be rotated such that the optimal fit between them constitutes one dimension of the rotated space. In that case, this dimension now contains information about all of the source dimensions that load onto it, and is thus likely to be more informative about the total structure of the data than any of them individually.

Directly operating on the data matrix usually has beneficial aspects. Crucially, it is comparatively easier to map the results back to the individual features that they originate from – if the aggregation determines a pattern, it is possible to see precisely which dimensions contribute to that pattern, aiding both interpretation and validation. Such information is much harder to retrieve from methods that require an explicit aggregational step. These, however, come with considerable advantages of their own. By abstracting away from the individual measurements, it is possible to find patterns beyond simple correlation, often resulting in a much better representation of the data as a whole. Furthermore, explicit aggregation enables a large variety of powerful methods, allowing the analyst to investigate research questions that would be difficult or impossible to tackle otherwise. The next section will detail the particularities of how such a step can be implemented.

2.1.1. Calculating linguistic distance

As noted in the previous section, many aggregational analysis techniques rely on the measurement of similarity or dissimilarity between taxa. Of these, dissimilarity measures are used more frequently, as they tend to be easier to determine.

Dissimilarity measures are a subset of distance measures. To qualify as a valid distance measure, the calculated values need to satisfy the following conditions, where $d(i, j)$ refers to the distance between taxon i and taxon j:

$$\mathrm{d}(x, y) \geq 0$$

$$\mathrm{d}(x, y) = 0 \Leftrightarrow x = y$$

$$\mathrm{d}(x, y) = d(y, x)$$

$$\mathrm{d}(x, z) \leq d(x, y) + d(y, z)$$

These conditions mean that all distances need to be non-negative, that the distance between two taxa is zero in exactly those cases when the taxa are equal, that the distance between two points is symmetrical, and that the distance between two points is not larger than the sum of distances between these two points and a third point. For dissimilarity measures, the last restriction is removed. As all methods discussed here satisfy this criterion, I will use the terms interchangeably.

Some techniques make use of similarity measures instead. In principle, similarities can be easily transformed into dissimilarities by using an appropriate function, such as the reciprocal:

$$\text{similarity}(i, j) = 1/d(i, j)$$

This transformation would satisfy the requirement that the more dissimilar two taxa are in any given measure, the more similar they are in the similarity measure. This changes the scale in a non-linear way, however, and similarities and dissimilarities would in general only be mildly negatively correlated. It is possible to find a better solution in those cases where a maximal dissimilarity can be defined in a meaningful way. This allows *normalization* of the distances, i.e. scaling the distances into the interval from 0 to 1, and subsequent conversion to similarities as follows:

$$\mathrm{d}_{\text{normalized}}(i, j) = \frac{\mathrm{d}(i, j)}{\max(d)} \text{ where } d \text{ is the set of all distances}$$

$$\text{similarity}(i, j) = 1 - \mathrm{d}_{\text{normalized}}(i, j)$$

Defining a meaningful maximal similarity is usually not problematic for categorical data, but can be difficult on other data types, especially for frequencies. In such cases, the researcher may choose an arbitrary number as the maximal distance that is at least as great as the maximal observed distance; the numerical values that result from this, however, do not necessarily have a straightforward interpretation.

One of the simplest categorical distance measures is the Hamming measure (Hamming 1950), which operates on strings of equal size. Measurements of taxa along categorical feature dimensions can easily be converted to such strings by mapping each categorical level per dimension to one character, then joining these in a fixed order. As an example, consider the binary data in Table 2.1. A binary distinction along 5 dimensions for two taxa has been mapped to 1 and 0 in each case, leading to the strings `00001` for taxon x and `10110` for taxon y. To calculate the distance between these strings, each position is considered individually, and the distance is increased by 1 if the characters at that position are not equal. In other words, Hamming distance counts the number of positions where the character needs to be substituted to change one string into the other. In the example, the only match is in the second of five characters; it follows that the Hamming

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
Taxon x	0	0	0	0	1
Taxon y	1	0	1	1	0
Distance	1	0	1	1	1

Table 2.1.: Example data illustrating the Hamming distance measure

distance is 4. Clearly, the maximal distance possible in the example is 5, leading to a normalized Hamming distance of 0.8, and an equivalent similarity of 0.2.

Hamming distance weighs each dimension and feature level equally. In the general case, this is a beneficial property. The analyst may nevertheless choose to use a different measure depending on the specific data and research question. For example, for binary data representing presence or absence of a feature, shared absences may not be considered interesting. In such cases, a metric like the Jaccard distance is more appropriate. It is defined as

$$1 - \frac{n_{\text{shared}}}{n_{\text{shared}} + n_{\text{different}}}$$

where n_{shared} is the number of shared presences and $n_{\text{different}}$ the number of differences. In other cases, the researcher may wish to place particular emphasis on rare feature values, such that taxa sharing a category in a given dimension are considered more similar when fewer other taxa share that category. Goebl (1984: I, 83–86) describes such a measure, the *Gewichteter Identitätswert*[1] $\mathrm{GIW}(x)_{jk}$. Categories that can be sensibly mapped to numbers – such as binary categories which can be mapped to zero and one – may also be measured using a numerical distance measure, to which we now turn.

Numerical data matrices can easily be interpreted as a real-valued space $\mathbb{R}^p$. Two distance measures in such a space are especially natural: the Euclidean and Manhattan distances. Euclidean distance is the length of the straight line connecting two points, and is defined as:

$$\mathrm{d}(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Szmrecsanyi (2011: 54) recommends this measure in the general case, as it is "well-known and fairly straightforward". The Manhattan distance, also known as city block distance,

[1] English publications tend to use Goebl's original German term, while Goebl (e.g. 2010: 444) himself uses *weighted identity value*, $\mathrm{WIV}(x)_{jk}$. The term *Gewichtender Identitätswert* is also used (e.g. Goebl 2007b: 199). The amount of weighting can be controlled through the parameter x, with $x = 1$ being the most common.

uses a rectangular line along each individual dimension, and is defined as:

$$\mathrm{d}(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_p - y_p| = \sum_{i=1}^{p} |x_i - y_i|$$

The advantage of this metric is that its interpretation is straightforward: it is the sum of all individual differences. Furthermore, it is identical to the Hamming distance when the data comprises only ones and zeros. The Manhattan metric also allows feature dimensions to be easily weighted, so that the contribution of individual dimensions to the final result can be adjusted.

The adequate measurement of the distance between linguistically meaningful strings is not trivial. While in some cases Hamming distance could be used, this is not satisfactory. I will detail the problems with such an approach using an example from Heeringa (2004: 122f.). In Savannah (Georgia), *afternoon* is realized as [ˈæəftəˌnʉn], and in Lancaster (Pennsylvania) as [ˌæftərˈnun]. In this case, the difference could thus be measured using Hamming distance, resulting in the following when stress is ignored:

æ	ə	f	t	ə	n	ʉ	n
æ	f	t	ə	r	n	u	n
0	1	1	1	1	0	1	0

This approach results in a distance of 5. It is immediately obvious that this is not an informative result: neither f, t, nor r are actually different, only their relative position in the string has changed. Furthermore, the distance could not be calculated if the strings were not of the same length, for example if the two dialects only differed in rhoticity. One way to solve this is through using Levenshtein distance (Kessler 1995). In addition to the substitution operation of Hamming distance, it allows two additional operations: deletion and insertion of characters. The distance is then the shortest sequence of these operations that converts one string into the other. The following shows one such solution:

æəftənʉn	Savannah, GA	delete ə	1
æftənʉn		insert r	1
æftərnʉn		substitute ʉ/u	1
æftərnun	Lancaster, PA		

The Levenshtein distance of these two strings is accordingly 3. This is still a single-feature distance measure, but an aggregate measure can be derived by combining the distances of all dimensions, for example by taking the arithmetic mean as in Heeringa (2004).

Measures of linguistic distance can be used with little further processing to tackle a number of research questions, such as the relation of geographic and linguistic distance, which will be discussed in Section 2.2.1. The next section will introduce two basic techniques in analyzing distance matrices, namely multi-dimensional scaling and clustering.

2.1.2. Fundamental distance analysis methods

Once the analyst has derived a distance matrix from the original data set, a wealth of analysis and visualization techniques become available. One approach involves finding a lower-dimensional arrangement that matches the original distribution of measuring points as closely as possible. This can be achieved by using the multi-dimensional scaling (MDS) family of procedures. I will illustrate how such an analysis is performed using a small, randomly generated data set.

Table 2.2 contains a randomly generated data set, in which six taxa are measured along five dimensions using a numeric scale ranging from 1 to 0. Using the method outlined in the previous section with a Euclidean distance measure results in the distance matrix depicted in Table 2.3. Now MDS can be used to find a configuration of points in a k-dimensional coordinate system that maintains the original distances from the distance matrix as accurately as possible; n is smaller than the original dimensionality p, with $k = 2$ being especially suitable for visualization on paper. Figure 2.1 shows a scatter-plot using metric two-dimensional MDS. This analysis reveals that taxa D, B, and F are rather close to one another, and the rest of the taxa are rather far away both from this group and from each other. How good is this representation? To determine this, the analyst can compute the correlation between the original and the new distances using Pearson's product-moment correlation coefficient r. This coefficient, when squared, indicates the proportion of the variance that the new set of coordinates can account for. In the depicted two-dimensional solution, r^2 is 0.8, indicating a good, but not perfect match. And indeed, the comparison of the distances in the scatter-plot with those in Table 2.3 reveals that, while the general pattern holds, the depicted distances within the D-B-F group are somewhat too small and others are too large. For example, $d(\mathrm{B}, \mathrm{F})$ is 0.71 and $d(\mathrm{B}, \mathrm{E})$ only slightly larger at 0.85. In the scatter-plot, however, the distance between B and F is clearly much smaller than between B and E. Adding a third dimension increases the correlation almost to a perfect match ($r^2 = 0.995$), a considerable improvement. This confirms that a two-dimensional solution is not completely adequate.

Another central technique concerns classification, or more precisely the identification of groups, subgroups, and their relations. One family of methods suitable for this purpose is cluster analysis. Typically, a cluster analysis builds a hierarchical classification in an agglomerative or bottom-up fashion, but variants exist that divide top-down, or that are not strictly hierarchical, such as fuzzy clustering or the network-based methods that will

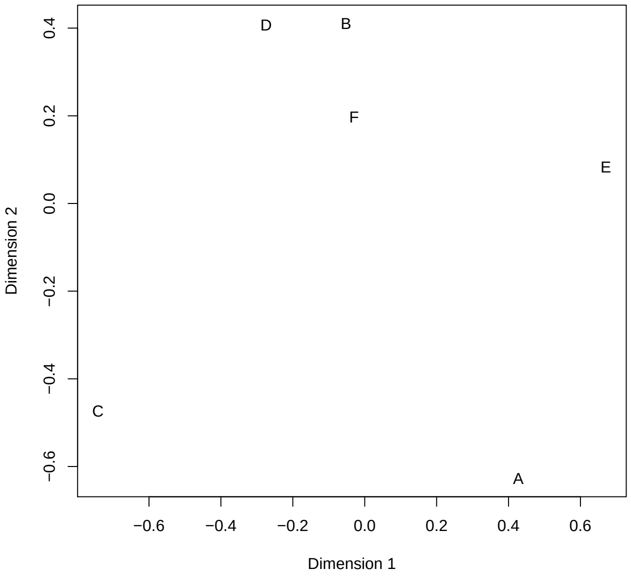


Figure 2.1.: Two-dimensional MDS analysis of example data set. Distances in the six-dimensional space are scaled to two dimensions.

be discussed later. The general manner by which a cluster analysis is performed follows these steps:

1. consider each taxon to be its own cluster
2. identify the two clusters x and y that have the shortest distance between each other
3. replace x and y with a new cluster representing the combination of both and recalculate the distances
4. repeat from step 2, until there is only one cluster left
5. the history of replacements now constitutes a hierarchical organization of the original taxa

To illustrate this, and thus make the interpretation of clustering results more accessible, let us return to the example distance matrix in Table 2.3.

The two taxa with the shortest distance are clearly B and D with a distance of 0.48. These taxa are removed from the distance matrix, and a new one containing the (B, D) cluster is inserted. Now, new distances between this cluster and the others need

	dimension1	dimension2	dimension3	dimension4	dimension5	dimension6
A	0.87	0.83	0.48	0.84	0.44	0.14
B	0.12	0.79	0.50	0.25	0.31	0.80
C	0.09	0.17	0.84	0.87	0.88	0.49
D	0.10	0.58	0.84	0.40	0.10	0.89
E	0.83	0.91	0.14	0.49	0.16	0.76
F	0.11	0.42	0.09	0.21	0.35	0.34

Table 2.2.: Example data set: data matrix. Six objects vary in six numeric dimensions.

	A	B	C	D	E	F
A	0.00					
B	1.17	0.00				
C	1.22	1.14	0.00			
D	1.29	0.48	1.08	0.00		
E	0.84	0.85	1.52	1.08	0.00	
F	1.16	0.71	1.17	0.99	1.02	0.00

Table 2.3.: Example data set: distance matrix. Euclidean distances between points in six-dimensional space.

to be calculated. The precise manner by which this is accomplished is one of the major variables distinguishing different cluster algorithms. For this example, a simple process called "single linkage" will be used, in which the new distance between two clusters is the minimal distance between any point from the first cluster and any from the second. For all distances except that to C, the shortest distance is that involving B. This results in the distances shown in Table 2.4. Now, the shortest distance lies between (B, D) and F. The cluster (B, D) is closer to all other points except A, leading to the distances in Table 2.5. Now, the shortest distance is 0.84 between A and E, which are fused next, resulting in the distances $d((A, E),((B, D), F)) = 0.85$ and $d((A, E), C) = 1.22$. In the next step, (A, E) and $((B, D), F)$ are fused, leaving only two taxa which automatically have the shortest distance. The full hierarchy is thus $((((B, D), F),(A, E)), C)$, which can be depicted as a tree diagram, or *dendrogram*, as shown in Figure 2.2. The height indicator at the right side of the plot indicates the distance at which a given cluster is merged; this is referred to as the *cophenetic distance.*

Figure 2.3 shows the result of other linking procedures. "Complete linkage", Figure (a), uses the furthest distance between members of different clusters as their distance. For this example, the main difference concerns the position at which taxon C is first merged into

	A	(B, D)	C	E	F
A	0.00				
(B,D)	1.17	0.00			
C	1.22	1.08	0.00		
E	0.84	0.85	1.52	0.00	
F	1.16	0.71	1.17	1.02	0.00

Table 2.4.: Example data set: distance matrix after the first iteration of the *Single linkage* hierarchical clustering algorithm.

	A	((B, D), F)	C	E
A	0.00			
((B,D), F)	1.16	0.00		
C	1.22	1.08	0.00	
E	0.84	0.85	1.52	0.00

Table 2.5.: Example data set: distance matrix after the second iteration of the *Single linkage* hierarchical clustering algorithm.

another cluster. In Figures (b) and (c), two ways of using averaged distances are shown. In the first of these, "average linkage" or, especially in bioinformatic contexts, "unweighted pair group method with averaging" (UPGMA), the average is scaled by the number of taxa in each cluster, whereas the second, "McQuitty's method" or "weighted pair group method with averaging" (WPGMA) uses the direct average and thus leads to a weighted result. In this specific case, both lead to a grouping that is essentially equivalent to single linkage, but with differing cophenetic distances. The final method, "Ward's method" presented in 2.3d, calculates the distance as proportional to the increase in variance that would result from merging those clusters. This method is widely used in dialectometric analyses (e.g. Heeringa & Nerbonne 2001, Szmrecsanyi 2011) and tends to result in "compact, spherical clusters" (R Development Core Team 2010: `hclust`).

One final clustering approach will be described here, phenogram construction as implemented in the neighbor-joining (NJ) algorithm (Saitou & Nei 1987, Studier & Keppler 1988), as well as one powerful extension to phenogram trees that removes the restriction of strict hierarchy. These methods originate in the bioinformatic reconstruction of evolutionary history, or phylogenetic inference, that will be discussed in its relation to historical linguistics in Section 2.2.2. The principles behind the view on clustering employed in these techniques has found followers beyond the purely historical domain. The first major difference results from the fact that these methods result in a dendrogram that does not

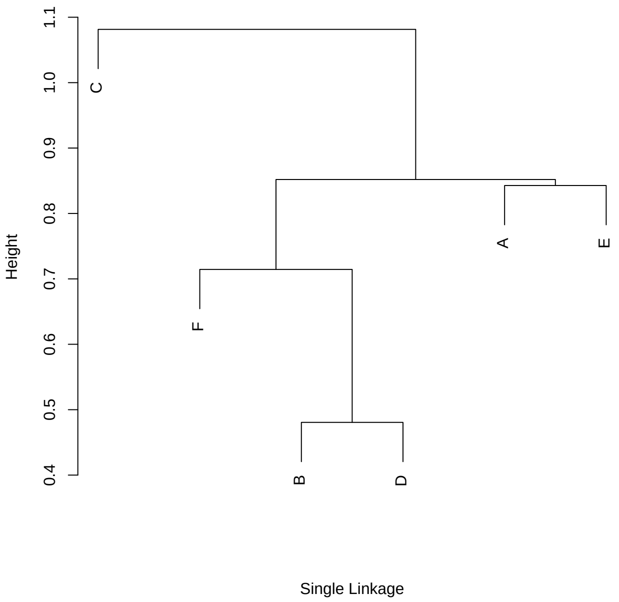


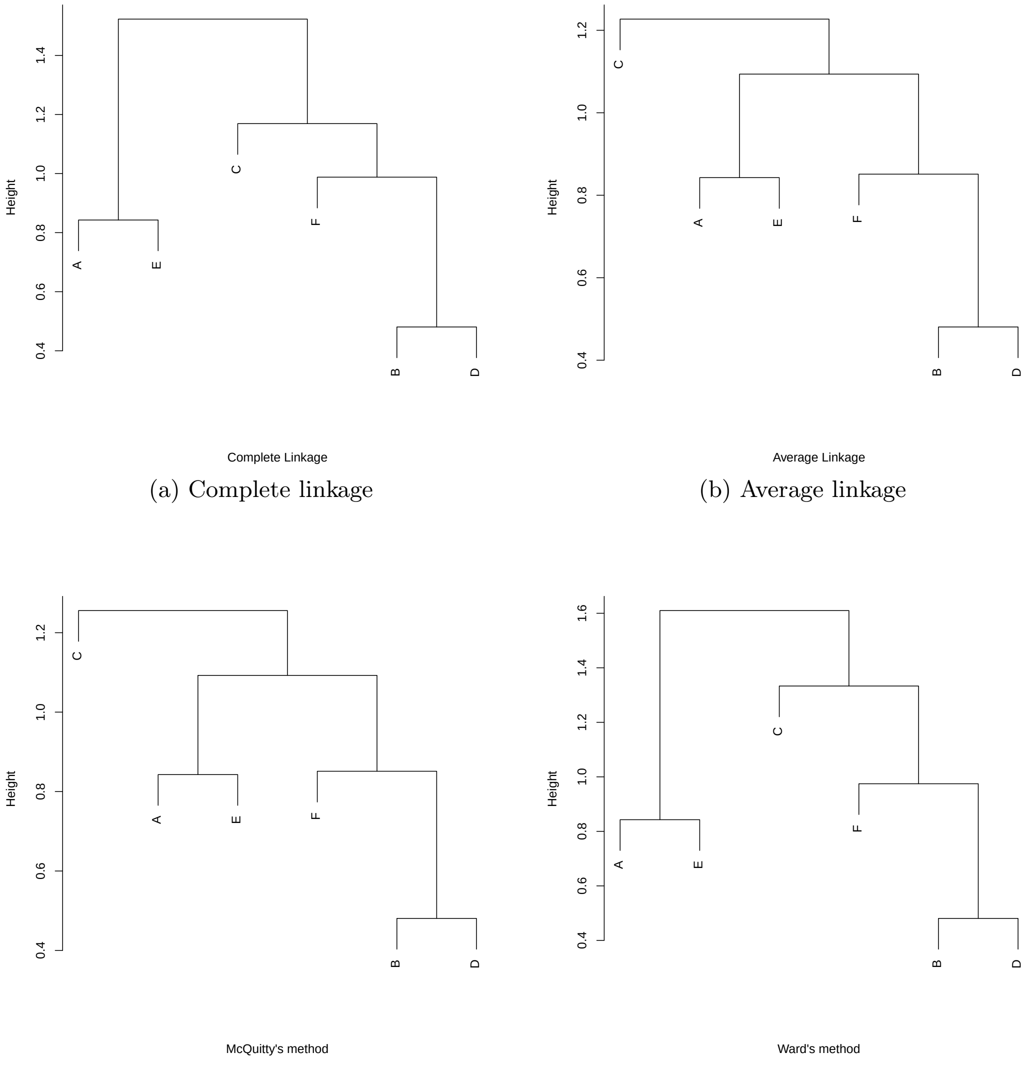
Figure 2.2.: Clustering result: single linkage. Dendrogram displays the hierarchical order in which clusters are fused.

have an explicit root. Rooted trees, by virtue of possessing a special node that represents the origin, have an up-down orientation; in dendrograms resulting from hierarchical clustering, this root is not always explicit, but is implicitly placed at the final clustering step, as evidenced by the evident directionality of the representation. Unrooted trees lack such ordering and thus there is no ancestor/descendant relationship. Instead, connections between nodes are to be understood as *splits*, bisections of the data into two groups, and trees are collections of *compatible* splits[2]. Rooted trees can be converted into unrooted ones by replacing the root and the branches extending from it with a single split, and unrooted trees can be transformed to rooted trees by choosing a split and inserting a root node in their middle (Dunn et al. 2008: 723, see also their Figure 1 for a visualization of the process). Second, splits may have differing lengths, and the length of the split between two nodes is a representation of their distance. Such trees, called phenograms, can thus be viewed as accurate a depiction of the underlying distance matrix as is possible in a compatible split system. Many methods use a least-squares estimator (or an equivalent

[2] A split system Σ is compatible, "if, for any two splits $S_1 = \{A_1, A_1'\}, S_2 = \{A_2, A_2'\}$ in Σ, one of the four intersections

$$A_1 \cap A_2, A_1 \cap A_2', A_1' \cap A_2 \texttt{or} A_1' \cap A_2'$$

is empty" (Huson 1998: 68)



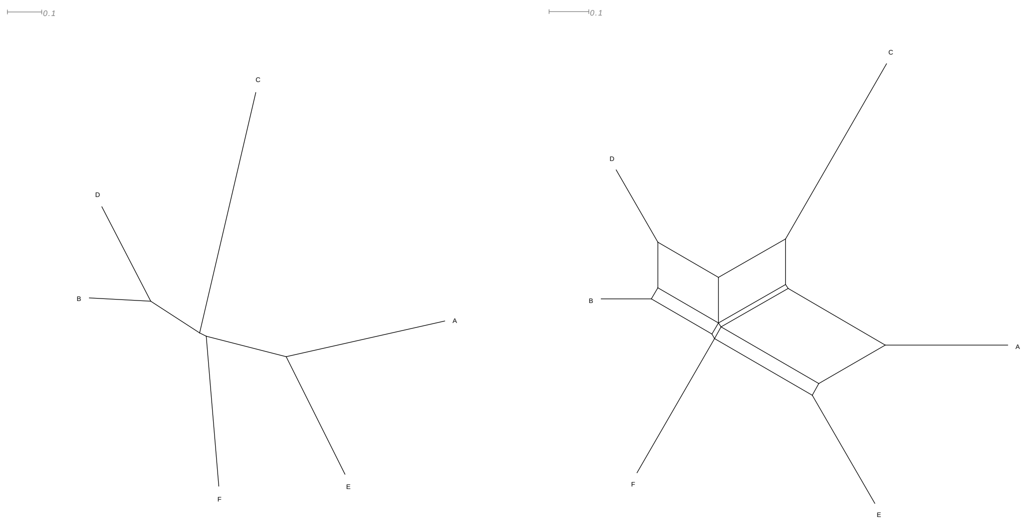
(a) Complete linkage

(b) Average linkage

(c) McQuitty's method

(d) Ward's method

Figure 2.3.: Clustering solutions using other linking algorithms. Dendrograms display the hierarchical order in which clusters are fused.



(a) Neighbor Joining

(b) NeighborNet

Figure 2.4.: Phenograms for the example data set.

process) to determine branch length (e.g. Saitou & Nei 1987: Appendix 1).

The NJ procedure bears similarity to the general clustering procedure outlined above, but is conceptually somewhat different. NJ starts out from a star-like constellation, in which each taxon is joined to a single central node. Then, for each node pair, it is evaluated by how much the total sum of branch lengths can be reduced by introducing a new node between these nodes and the central node. This results in a new distance-like matrix, from which the best value is chosen, and the respective node is inserted; let us assume this node is connected to taxa A and B and is labeled u_1. Now, both the branch lengths for the connections from A and B to u_1 can be identified, and we can replace A and B in the original distance matrix with u_1 by averaging over the remaining differences (i.e. after subtracting the already established branch lengths). This process can then iteratively be applied on the resulting distance matrix, until all branch lengths are determined (see Saitou & Nei 1987: Figure 3 for a visual example of the process).

Figure 2.4a shows the result of this algorithm on the example data. As with the other clustering algorithms, the groupings (A, E) and (D, F) are made; the other taxa are relatively unconnected to either group. This solution has a good Least Squares Fit of 99.18, as calculated by the phylogenetic software package *SplitsTree* (Huson & Bryant 2006).

When considering data in terms of splits, it is not difficult to conceive of cases where

	Feature	O&S	ScE	IrE	North
[1]	*them* instead of demonstrative *those*	0	1	1	1
[5]	Object pronoun forms serving as base for reflexives	1	0	1	1
[33]	*after*-Perfect	0	0	1	0
[61]	relative particle *what*	0	0	0	1
[2]	*me* instead of possessive *my*	0	0	1	1
[39]	levelling of preterite/ppt verb forms: part. replacing the past form	0	0	1	1
[7]	*she/her* used for inanimate referents	0	1	0	1

Table 2.6.: Data used for the demonstration of network diagrams. Seven features out of the morphosyntactic part of the *Handbook of Varieties of English* survey, for four different varieties. From Wolk (2009: Table 2.4).

the splits are not compatible. Consider the following example from Wolk (2009). The data set used here is a subset of the morphosyntactic survey of the *Handbook of Varieties of English* (Kortmann et al. 2004) and can be found in Table 2.6. It consists of binary presence information on six morphosyntactic features in four British varieties: Orkney & Shetlands English (O&S), Scottish English (ScE), Irish English (IrE), and the dialect of the North of England (North). Absence is coded as zero and presence as one. As this is categorical data, the Hamming distance as described in the previous section is an adequate choice.

When only the first four features in Table 2.6 are taken into account, the NJ algorithm leads to the simple classification shown in Figure 2.5a. Each variety differs from the others in exactly one characteristic and no meaningful grouping can be made, yielding a star-like shape. All varieties are equally different from one another. The next two features in Table 2.6 change this: both are present only in the North and IrE, which suggests a split with these dialects on the same side. This is visualized in Figure 2.5b. Both of these trees perfectly represent the underlying data.

Adding the final feature from Table 2.6, [7], however, is not straightforward. This feature would induce a split that groups O&S with IrE and ScE with the North, but this split is not compatible with the existing grouping of ScE and O&S. Patterns like this can not be adequately represented as a tree, as the underlying signal is not strictly hierarchical. Forcing it into a dendrogram means that part of the information will be lost.

Methods have been developed that allow the visualization of such non-compatible split systems as *splits graphs* (Dress & Huson 2004). Instead of using simple branches to represent a split, sets of parallel lines are used, resulting in network diagrams instead of

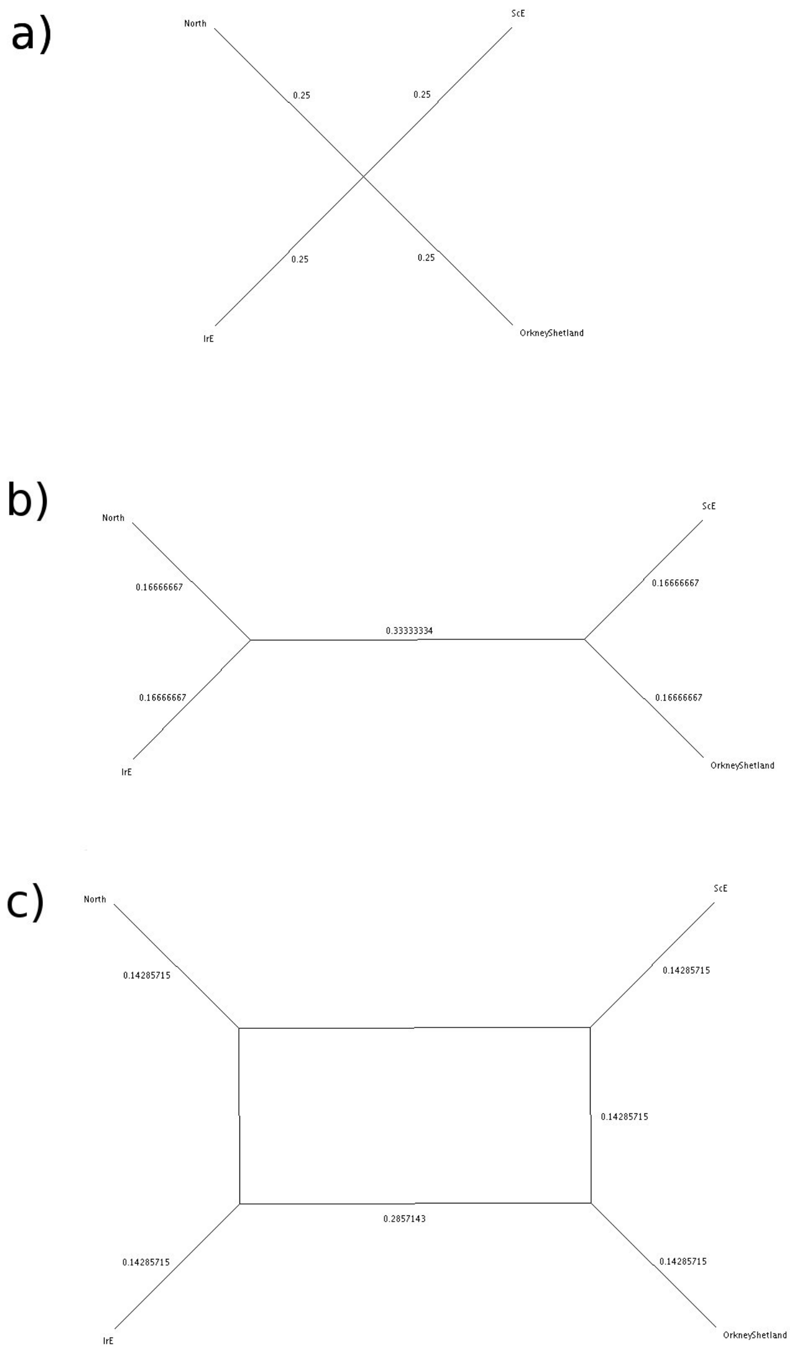


Figure 2.5.: Example NJ dendrograms and NeighborNet splits graph using the data from Table 2.6. a) Features [1], [5], [33], and [61]. b) Features [2] and [39] added. c) Feature [7] added. From Wolk (2009: Figure 2.5).

trees. To reduce visual complexity, split systems that can be displayed as planar graphs, i.e. without crossing lines, are especially beneficial. One popular algorithm for finding such a system from a distance matrix is an extension of NJ called NeighborNet (Bryant & Moulton 2004, Huson & Bryant 2006). It proceeds in a similar fashion and uses the same criterion, but when a relation between two points is identified, the split is not inserted immediately. Instead, the two points are just marked, and the procedure is repeated until the same point is marked twice. Then, two splits are inserted, each representing the doubly marked point in relation to one of its marked neighbors. This process is repeated until only three clusters are left. The fusion sequence can subsequently be used to generate a network-like diagram. A beneficial aspect of this procedure is that the results will not be needlessly complex: for cases where a segment of the data can be adequately represented as a hierarchical tree, the corresponding segment of the network will be tree-shaped.

Figure 2.5c shows the results of using this algorithm on the full data from Table 2.6. The split that groups IrE with the North is still present, but the incompatible split is now also visible. Together, they form a boxy shape, referred to as a *reticulation*. The resulting network diagram can represent the underlying distances perfectly. For data sets with more than four elements and complex signals, this is not necessarily true, as NeighborNet enforces planarity. In general, however, allowing reticulations leads to a notable improvement compared to the corresponding tree.

Let us now return one final time to the data set in Table 2.2. Figure 2.4b shows the splits graph that the NeighborNet algorithm yields. The least squares fit of this network is very good (99.85), indicating that it is a very accurate description of the underlying distances. This solution is not tree-like, and it has no tree-shaped segments. Instead, we can find pairwise similarities linking B to A (through D and C). There are also two splits that group the taxa into two equally-sized groups, one which contains (B, C, D) and one with (B, E, F). Interestingly, the (B, D, F) grouping that appeared in the hierarchical clustering dendrograms in Figures 2.2 and 2.3 does not appear in the network, although such a split could have been inserted. This indicates that, while these three points are close together, positing them as a group does not improve the representation, and their relation is thus of limited explanatory value. I will return to phenograms and networks in Section 2.2.2.

One final point can be made on the basis of this example. Each technique described here focused on another particular aspect of the data. While the results shared many similarities, there were also crucial differences. This is true of aggregational approaches in general. Furthermore, individual methods may be very sensitive to small changes in the data (see e.g. Nerbonne et al. 2008). Thus, any result should be carefully considered in light of the several perspectives that are available, and techniques for testing and increasing the robustness of the results should be used where feasible.

2.2. Applications of aggregational techniques

2.2.1. Dialectometry

Dialectometry (DM), "the measuring of dialects", is a research paradigm in dialectology. Its central insight results from the observation that geolinguistic investigation through single features is practically guaranteed to result in a noisy and/or inaccurate picture of the overall reality. This noisy picture is likely incompatible with both other single features and with speakers' perceptions of how the dialects are related to each other. The best way of finding the true geographic signal, then, is not to simplify the data for a single feature through qualitative abstraction, as is done in the interpretative maps of classical dialectology (Chambers & Trudgill 1998), but to combine a large number of features into a single aggregate analysis. A frequently used metaphor is that of finding the "forest behind the variable trees" (Spruit et al. 2009: 1642); regardless of the precise nature of distortions on a single 'tree', aggregation allows the individual noise factors to cancel each other out, and thus leads to greater accuracy in representing the true signal of the multidimensional 'forest' (cf. also Szmrecsanyi 2008: Section 2).

The field was pioneered and named by the French geolinguist Jean Séguy, the director of the *Atlas linguistique de la Gascogne* project, at the beginning of the 1970s. While the primary result of that project was a traditional dialect atlas containing single-feature maps, the appendix to the sixth and final volume held the first aggregated maps for the different linguistic levels contained in the atlas. On these maps, the lines between measuring points contained a simple measure indicating the percentage of disagreements between the locations (cf. Chambers & Trudgill 1998: 137ff.; Heeringa & Nerbonne 2013).

The manner of aggregation and visual presentation were still quite crude, yet the general idea proved very influential. The first research group that refined these methods was founded by Hans Goebl (1982, 1984, 2006, 2010) at the end of the 1970s and is now commonly referred to as the Salzburg school of dialectometry (Salzburg DM). Goebl's work formalized the approach used by Séguy and extended both the methodological apparatus and the types of visual presentation considerably. The methods of the Salzburg school are rooted in taxonomy, and are intended to foster a "qualitative geolinguistics via quantitative means" (Goebl 2010: 436) with the primary goal of uncovering

> [t]he effects of the dialectal and basilectal management of geographic space

> by Homo loquens that is executed according to determinant communicative, social and other similar principles. (ibid.)

The first step of analysis in this paradigm involves the extraction of nominal values from the individual maps of a dialect atlas, often splitting the contents of a single source map onto several categories, a process labeled "taxatation". This selection and classification step is a thoroughly qualitative one, building on the analyst's existing knowledge to carefully identify and categorize the linguistically interesting features from the raw atlas data. After an aggregation process[3] as outlined in Section 2.1.1, many different results can be extracted. As a geolinguistic enterprise, Salzburg DM typically relies on visualizing them in a large number of different maps. Typical applications include the display of the relations between neighboring dialects, group identification, and analysis of the homogeneity within groups. Many of these visualizations are based on a polygonization of the area under study using *Voronoi tesselation* (Goebl 2006: 417 & Figure 3). The resulting polygons can then be colored to highlight different aspects of the aggregational result, resulting in so-called *choropleth maps*. There are many types of properties that can be visualized in this way, including similarities to a reference variety (ibid.: 418 & Maps 3–6), summary statistics of the distribution of linguistic similarities at each location such as the maximum or the skewness (ibid.: 419f. & Maps 7–8), the results of cluster analyses (ibid.: 420f. & Map 9) or correlations to language-internal or -external characteristics (ibid.: 421f. & Maps 13–16). Beyond choropleth maps, Salzburg DM utilizes interpoint maps, in which either the sides of triangles, which are derived from a triangulation of the area under study, are colored to indicate the similarity ("beam maps") of a measuring point to its immediate neighbors, or the sides of the corresponding polygons are colored to indicate their dissimilarity ("honeycomb maps") (Goebl 2010: 447ff. & Maps 2209–2210). Salzburg-style analyses have focused on romance dialects in France (see e.g. Goebl 2006: references on p. 415) and Italy (Goebl 2007b), but were also applied to other languages such as Catalan (Rivadeneira & Casassas 2009) and English (Goebl & Schiltz 1997).

The second influential school in the development of DM, the Groningen-based group led by John Nerbonne (e.g. Nerbonne 2009, Heeringa 2004), emerged during the later half of the 1990s. Much of their work eschews the taxonomic categorization of the primary data that is characteristic of the Salzburg method. Instead, their methods operate directly on the character strings that are naturally extracted from lexical or pronunciation data, using variants of the Levenshtein algorithm discussed in Section 2.1.1. The Groningen school is characterized by a considerable methodological awareness, carefully evaluating the effect of different ways of conducting measurements and analyses. Crucial in their work is the

[3] Salzburg DM typically relies on similarities instead of distances

comparison of aggregational results to external factors for validation, for example via experimentally determined perception judgments (Heeringa 2004: chapter 7). They also use their results to test external hypotheses, for example the question to which extent surnames behave like lexical items and the impact of that on using surnames as a proxy for genetic variation (Manni et al. 2008). A central topic is the precise nature of the relation between geographic and linguistic distance. This relation is used in two ways. First, it is studied in itself (Nerbonne 2013), confirming earlier dialectometric findings of a sublinear relationship (Séguy 1971, Goebl 2005) and rejecting Trudgill's linguistic gravity hypothesis (1974) which posits a quadratic relationship (Nerbonne & Heeringa 2007). Second, it is proposed as an evaluation criterion, allowing researchers to choose between different ways of performing aggregation: per the *Fundamental Dialectological Postulate* (Nerbonne & Kleiweg 2007), "[g]eographically proximate varieties tend to be more similar than distant ones", and thus a good method of aggregation should generally result in greater local coherence than a bad one.

Some visualization methods that the Groningen school of DM uses are similar to those of Salzburg DM. For example, choropleth maps are commonly used to depict the results of hierarchical clustering and are sometimes referred to as *color area maps* (e.g. Heeringa 2004: 164). The school has also contributed some new methods for cartographic mapping to the dialectometric apparatus. One is the *composite cluster map* (Nerbonne et al. 2008: Fig. 4). It is related to the honeycomb map, but instead of coloring by the interpoint distance, the thickness of the polygon lines is scaled to their cophenetic distance in a hierarchical cluster analysis. The composite cluster map thus provides a gradient representation of dialect areas. The second type of map is named *continuum map*. It projects the results of a three-dimensional MDS analysis onto a cartographic representation by linking each dimension to one axis of the RGB (red, green, blue) color space, and thus generating a unique color for each location (Heeringa 2004: 161ff.). Intermediate points can then be assigned a color through interpolation. Continuum maps yield a direct representation of the gradience in a data set.

Within the last few years, two new directions in dialectometric research have emerged. The first is the introduction of explicit models into dialectometry. There are two variants to this. The first was pioneered by Martijn Wieling and involves using generalized additive models to evaluate how much the dialects spoken at individual locations differ from the standard variety, typically evaluated using Levenshtein distance. This approach has been used on Dutch pronunciation (Wieling et al. 2011), Catalan pronunciation (Wieling 2012) and Tuscan lexicology (Wieling et al. forthcoming). In contrast to most other approaches in dialectometry, the distances between dialects are not evaluated, only the distance to the standard. The second model-based approach was developed at the universities of Augsburg and Ulm under the supervision of Werner König and Stephan Elspaß. It moves the focus from aggregated analyses back to individual features; in their case lexical (later also morphological and phonetic) variation in southwestern Bavaria using the *Sprachatlas von Bayerisch-Schwaben* (König 1997 – 2009). They use a mathematical technique called *intensity estimation* to estimate the probability that a certain lexical variant will be used at a given location. This is intended to account for sampling variance: "a single record at a single site that was uttered by an informant in a specific interview situation may or may not reflect common usage in the local dialect" (Pickl et al. 2014: 25). The probabilities can then be projected onto so-called area class maps, indicating the gradient dominance and competition of variants (Rumpf et al. 2009). The results can then be analysed using variants of cluster analysis to determine large-scale spatial patterns in the lexical realizations of concepts (Rumpf et al. 2010, Pröll 2013). Recent work has begun to integrate this with more traditional aggregational metrics such as linguistic distance (Pickl et al. 2014), but again centered around improving individual maps. Linguistic distance is here not the result, but an intermediary step leading to better areal predictions for individual realizations.

A second recent strand in dialectometric research, the integration of dialectometric techniques with corpus-based frequency measurements, will be discussed in Section 2.3.

2.2.2. Aggregational techniques in historical linguistics

One of the pillars of historical linguistics is the study of developmental relations between languages. Its goal is a complete reconstruction of the family history of languages. In such a history, each language is linked to the language from which it descended, until all languages of one family are connected to their common ancestor genetically, i.e. through an unbroken series of speakers learning from one another. The major tool of this line of research is the so-called *comparative method*, which consists of several components that are applied in an iterative manner. First, connections between languages are established based on their observed features and the specific differences between them; these then allow the formulation of hypotheses on the rules that govern the changes occurring in that set. This knowledge then helps to refine the original observations by eliminating misclassifications that result from chance or non-genetic factors. These refinements may result in refined hypotheses, and finally in the reconstruction of unobserved common ancestor languages. The precise changes from the ancestor to the descendants allows the identification of fine-grained subfamilies. In general, this method is applied to lexicophonemic data, where cognates - words that are genetically related - as well as the sound changes that underlie the differences between realizations of the same cognate are identified.

The investigation of language family histories is an undertaking that requires an aggregate view on the data: the words corresponding to a single meaning may have noisy histories, and the real history will only become apparent when considering many meanings at the same time. Early applications, which usually had the Indo-European language family as their subject matter, were very successful, but their data collection methods were not amenable to quantitative investigation. A crucial development toward more formalization was the introduction of the lexicostatistical paradigm by Morris Swadesh (1950). Lexicostatistics replaces convenience samples of linguistic features with a carefully selected, fixed set. This set consists only of basic meanings, whose realizations are both likely to be present in each language and unlikely to result from non-genetic change, such as borrowings. Cognancy judgments are then performed on these items, using the standard tool set of the comparative method. These can then be used to calculate the proportion of shared cognates between two languages, serving as a measure of their relatedness. On the basis of this information, the branching order of the families can be determined. A sister discipline of lexicostatistics named glottochronology then attempts to place dates on this branching order using estimates of the lexical replacement rate. This method is controversial, and has largely fallen out of favor today (cf. McMahon & McMahon 2005). The general idea behind lexicostatistics, however, has received a surge in popularity due to the availability of electronic versions of Swadesh-type lists, for example the one used by Dyen et al. (1992), and crucially due to the availability of a large and varied methodological apparatus in another discipline: phylogenetic inference as used in bioinformatics.

Linguists often use evolution as a metaphor for language change. Consider, for example, the concept of genetic language change that was mentioned above, and the notion of "[t]he evolution of Postcolonial Englishes" found in Schneider (2007: title of Chapter 3). This metaphor is usually mapped in the following way: languages and language varieties correspond to species, speakers (and their linguistic knowledge) correspond to members of a species, and language features as represented in a speaker's linguistic knowledge correspond to the genetic information that a member of the species carries. Assuming these correspondences, the task of the historical linguist working on language families is very similar to that of the evolutionary biologist trying to determine the branching order in the tree of life. The discipline of bioinformatics has developed many methods for determining such phylogenies, and several of them have been successfully applied to linguistic data. The subject of most investigations so far is the Indo-European language family, for example in the work of the CPHL project (Computational Phylogenetics in

Historical Linguistics, e.g. Ringe et al. 2002), as the good data availability allows robust analyses, and the large amount of widely accepted prior research provides a backdrop for evaluation of different methodologies. More recent work has begun to extend its focus beyond Indo-European, such as Michael Dunn's work (e.g. Dunn et al. 2008) on the languages of Island Melanesia, or the Quantitative Historical Linguistics project[4] led by Michael Cysouw, which deals with the indigenous languages of South America. All these studies rely on categorical data as input and usually deal with lexical data, often supplemented by phonological or morphosyntactic information.

Two types of bioinformatic methods can be distinguished, phylogenetic and phenetic (see also Section 2.1.2) ones. A phylogenetic method leads to a result that constitutes an exact hypothesis about evolutionary history, while a phenetic method "works on the basis of observed similarities and distances between languages at a particular time, and does not explicitly seek to reconstruct a history for the group" (McMahon & McMahon 2005: 158). Phylogenetic methods often, but not always, directly operate on the observed categorical data ("character states"), reconstructing the hypothetical realizations of ancestor nodes. All possible trees are considered, and the ones corresponding best to a relevant criterion are selected. A rather straightforward criterion is Maximum Parsimony, the total amount of character state changes. In other words, when comparing two trees, the history that involves the fewest developments is best. The method used by Ringe et al. (2002: 73) is somewhat related, minimizing the number of characters that are *incompatible* with the tree, i.e. that do not map to exactly one connected sub-graph. More elaborate character-based methods exist as well, such as Bayesian phylogenetic inference, which is based on a formally specified probabilistic model of evolution that can accommodate prior knowledge about the likelihood of certain character state changes. Dunn et al. (2008) use such an approach. In general, one of the advantages of character-based methods is that it is possible to identify the exact meaning of each branching in the resulting tree in terms of what exactly the differences in the data between the two descendants are. The cost of this is that the methods also require rather clean and historically appropriate data. In contrast, some phenetic methods, and in particular those resulting in a splits graph representation like the NeighborNet algorithm (discussed in Section 2.1.2), have very wide ranges of application.

Splits graph representations, as phenetic networks, are an adequate choice when the evolutionary history is mixed through genetic recombination, a process in which strands of DNA or RNA exchange genetic material directly (Bryant & Moulton 2004). Such mixed histories also emerge in linguistic material, when speakers from different languages or

[4] http://quanthistling.info

dialects exchange characteristics through language contact, for example in the form of lexical borrowing. Thus, such methods are easily and flexibly applied to historical, historically minded dialect-phonological (McMahon et al. 2007) and typological (Albu 2006), as well as purely dialectometric (Szmrecsanyi & Wolk 2011) and typological (Cysouw 2007) purposes. Their allure is that they allow patterns in the data to emerge, without the assumption that the data necessarily fall into neat hierarchical groups. Messiness and intersections are allowed. Therefore, as Bickel (2012) notes, "[o]ne method from phylogenetics, split graphs, is a useful tool for similarity analysis, even without stakes in evolutionary explanations". In this spirit the NeighborNet algorithm will be used in Section 5.3.

In dialectometric analyses, use of these methods is still rare. This, as Prokić & Nerbonne (2013: 153) argue, is due to the fact that "there is no direct way to link this kind of representation and geographic data, i.e. to project data onto the map, which is very important element [sic] of the research in traditional dialectology and in dialectometry as well". They instead recommend MDS-based continuum maps, as introduced in the previous section. One clear advantage of splits graphs, however, is that the relations are visible directly, whereas on continuum maps they have to be inferred from color similarities. I will therefore present the two in combination.

2.3. Corpus-based dialectometry

A final, very recent addition to DM is the introduction of corpus-based investigation of dialect features and their frequencies. This line of research tends to focus on morphosyntax, although Szmrecsanyi (2013: 4) lists the frequency-based work on phonology by Hoppenbrouwers & Hoppenbrouwers (2001) as a predecessor in spirit. Three lines of research that are relevant here are Szmrecsanyi's "corpus-based dialectometry" (CBDM), Grieve's "multivariate spatial analysis", and the bottom-up approach by Sanders.

Szmrecsanyi's CBDM is the direct foundation for all the work presented here, and we will revisit its methodology and results throughout the following chapters. The discussion here will be kept high-level and brief. CBDM is, in short, an integration of the corpus-based inquiry of dialectal features (Anderwald & Szmrecsanyi 2009) with dialectometric techniques. The major motivation for this, in addition to the benefits of aggregational analysis in general, is the insight that "*compared to linguistic atlas material, corpora yield a more realistic linguistic signal*" (Szmrecsanyi 2013: 4). In contrast to what is essentially the judgments of individual informants and fieldworkers as they can be found in dialect atlases, using collections of text that occurred naturally as the primary data source yields several benefits. Whereas the former is "categorical [and] exhibits a high level of data reduction" (ibid.), investigations of naturalistic frequencies offer a much more gradient signal. Szmrecsanyi's method begins by extracting a catalog of features based on the existing literature on morphosyntactic variation in varieties of English, then conducts a corpus analysis on a suitable dialect corpus, in this case the *Freiburg Corpus of English Dialects* (Hernández 2006), counting occurrences of each feature (Szmrecsanyi 2011: 49ff.). Then, the counts undergo mathematical transformations to make them more suitable for statistical analysis. This crucially includes a normalization step, in which the frequency differences are scaled to a common number to make areas of different text size comparable. Whether this correction is adequate will be one of the central topics of this dissertation. Afterwards, the counts can be transformed into a distance matrix, and finally a variety of methods from the apparatus of the Salzburg and Groningen schools of dialectometry is applied on the results. Among the many results that are particular to Great Britain, Szmrecsanyi finds one with potentially far-reaching implications: compared to other research, the relation between geographic distance and language variability behaves very differently. In contrast to other studies, which largely found a strong correlation and a sub-linear relationship (see Section 2.2.1), Szmrecsanyi only finds a very weak association that is furthermore linear in nature. Szmrecsanyi argues that the most plausible explanation for this is that the categorical signal in other analyses is overly reduced, omitting crucial facts about actual linguistic diversity. This data reduction could be considered "essentially a form of academic fraud" (168).

Another approach to the combination of corpus investigation and DM is the one pioneered by Grieve (2009), "multivariate spatial analysis". It exhibits some similarities to Szmrecsanyi's method, but also crucial differences. Grieve is interested in studying the geographic distribution of grammatical variation in written Standard American English, and makes several innovations to this end. Grieve begins by collecting a corpus of letters to the editor in local American newspapers. As many newspapers have extensive online archives, he was able to collect a large amount of material, comprising in total 25 million words spread over 200 cities. He then, like Szmrecsanyi, compiles a feature set. This set contains 45 grammatical alternations where the probabilities for both realizations can be determined or approximated automatically using computer-linguistic methods. Grieve then processes the raw probabilities using two techniques. First, he determines both global and local spatial autocorrelation. Global spatial autocorrelation, more specifically Moran's I (Cliff & Ord 1973), measures the degree to which higher values tend to be closer to other higher values, and lower values closer to other lower values. All features where this value is not significant are removed from the analysis. Next, Grieve calculates the local spatial autocorrelation for each location and feature via Getis-Ord *Gi** (Ord & Getis 1995), a form of hot spot analysis. In this particular application, a 500 mile radius was placed on each location, and all locations in that radius were included as the point's neighbors. Then it is determined to what degree that point and its neighbors have either particularly high or low values. This is a form of geographical smoothing, as each location is pooled with its neighbors to determine the score. These values are then processed again with a principal components analysis, a technique that can reduce the dimensions of a data set by finding the common elements. The results of this are then used as the input to a hierarchical cluster analysis; only the six components that contribute the most to the overall pattern, which together account for 92 percent of the variance, are included. Using this method, Grieve is able to identify 12 geographically coherent dialect regions. In Szmrecsanyi's terms, however, Grieve makes use of three data reduction techniques: eliminating features without a spatial distribution, smoothing via *Gi**, and removing variation through principle components analysis.

Sanders (2007; 2010) uses a very different approach. He eschews constructed feature lists, as they are "subject to bias from the dialectologist" (2010: 5). Instead, his method attempts to leverage the automated tagging and syntactic parsing methods available in the tool set of computational linguistics to derive a measure of syntactic distance from the bottom up. To do so, he extends a method for identifying significant differences in part-of-speech trigram frequencies developed by Nerbonne & Wiersma (2006) to operate on syntactic trees, then develops that into several distance measures based on the derived counts and subjects them to dialectometric analysis. Sanders (2007), the first application of this method, concerned itself with British dialects using the *International Corpus of English, Great Britain* corpus (ICE-GB) as the data source. Sanders was able to find that some dialect areas were significantly different from others and identified a rough north/south axis, but did not perform full dialectometric analysis. His later application on a Swedish dialect corpus did use dialectometric visualization techniques, and showed that his approach is able to distinguish the major dialect areas despite a rather small corpus. Nevertheless, the match between geographic and linguistic distances is quite low.

In the next chapter, I will present my extensions to these methods. I will argue that the normalization step in CBDM alone is not enough to make subcorpora comparable, and replace it with smoothing that takes either the total characteristics of the distribution or the local and global geographic context into account. Finally, I will present my variant of Sanders's method, in which I add a method to assess a feature's distribution throughout the corpus by means of permutation.

3. Data and Methods

This chapter begins with an introduction of the data source tapped in this work, the *Freiburg Corpus of English Dialects*. Then, Szmrecsanyi's *corpus-based dialectometry* methodology will be discussed. Potential problems with this method will be demonstrated, and two ways of solving them will be introduced: *mixed-effect models* using lmer and *generalized additive models*. Then, I will detail a bottom-up approach that counts and analyzes part-of-speech co-occurrences. A brief summary will conclude this chapter.

3.1. Data

3.1.1. The Freiburg Corpus of English Dialects

All analyses that will be discussed in the later sections tap the Freiburg Corpus of English Dialects (FRED, cf. Hernández 2006) as their source of data on morphosyntactic variability in Britain. FRED is a spoken dialect corpus containing orthographic transcriptions of oral history interviews with speakers from multiple locations in England, Wales, Scotland and the Hebrides. In these interviews, the informants were asked about their "life memories" (Hernández 2006: 1). In total, the corpus contains about 2.5 million words spread over 372 interviews. Work on FRED began at the University of Freiburg in the year 2000 and proceeded as follows:

> Tape and mini-disc copies were made of pre-selected original tape recordings made available by various fieldworkers, historians, local museums, libraries and archives from different locations in England, Scotland, Wales, the Hebrides and the Isle of Man. Back at Freiburg University, the tapes were digitised for protection [...] and stored on DVD. The interviews deemed most suitable for our purposes were then transcribed (either from scratch or revised) by English native speakers and linguistically trained staff. (Hernández 2006: 2; footnotes removed)

Each interview is labeled with an identifier consisting of a three-letter county identifier and a natural number with up to two leading zeros, combined with an underscore. The county identifier is based on the Chapman county codes, "the standard format for genealogical purposes", for the administrative borders as they were structured before 1974 (Hernández 2006: 15). As an example, the first text in the Kent subcorpus has the identifier KEN_001.

The beginning of one corpus file can be found in (1). The file, labeled KEN_004, contains an interview by a single interviewer, IntMW, with a single male informant. This informant has the identifier KentPB and is a laborer from Tenterden in Kent. The interview was recorded in 1976, when KentPB was 87 years old.

(1) {<u IntMW> If you could tell me when you were born to start off?}
<u KentPB> #When I was born? #Well now look, you can get at it perhaps, I 'm, well we 'll say I 'm eighty-six and we 're nineteen now, seventy-six, idn't we. #How long would that be, eighteen and eighty-nine...?
{<u IntMW> Eighteen eighty-nine, eighteen ninety?}
<u KentPB> #Ay?
{<u Int> Eighteen eighty-nine?}
<u KentPB> #Eighty-nine, yeah, mm. #Well...
{<u IntMW> You were going to tell me what you could remember when you were a little boy?}
<u KentPB> #Ay?
{<u IntMW> You were going to tell me what, right back, what you could remember?}
<u KentPB> #Yeah. #Well now look, I was born at Benenden, Standen Street, in ehh, what 'd we say, eighteen eighty-nine?

The speaker of each utterance is indicated by the speaker tag (<u>) preceding it; parts contributed by the interviewers are identified by the presence of curly brackets ({}). A number sign (#) indicates the beginning of sentences.

One particularly relevant subset of FRED is the FRED Sampler (FRED-S, Szmrecsanyi & Hernández 2007), which only contains those texts where copyright restrictions do not apply. While FRED-S contains a much smaller amount of data – only about one million words from 144 dialect speakers – it is available with more extensive annotation, namely as a version including part-of-speech tags from the CLAWS7 tag set. (2) shows the beginning of the KEN_004 as it appears in this version of FRED-S:

(2) <u IntMW> If_CS you_PPY could_VM tell_VVI me_PPIO1 when_CS you_PPY were_VBDR born_VVN to_TO start_VVI off_RP ?_? </int>
<u KentPB> When_CS I_PPIS1 was_VBDZ born_VVN ?_? Well_RR

> now_RT look_VV0 ,_, you_PPY can_VM get_VVI at_II it_PPH1 perhaps_RR ,_, I_PPIS1 'm_VBM ,_, well_RR we_PPIS2 'll_VM say_VVI I_PPIS1 'm_VBM eighty-six_MC and_CC we_PPIS2 're_VBR nineteen_MC now_RT ,_, seventy-six_MC ,_, id_NN1 n't_XX we_PPIS2 ._.

Each word is annotated with a part-of-speech tag. For example, KentPB's first sentence begins with *when*, which is classified as `CS`, the tag for a subordinating conjunction. The following word, *I*, is tagged as `PPIS1`, the first person singular subjective personal pronoun. A list of all tags in the CLAWS7 tag set can be found in Appendix A.

For the analyses presented in the upcoming chapters, only subsets of the complete corpus were suitable. First, there is very little data available in some counties, making quantitative analysis impossible. Texts and counties were excluded when less than 5,000 words of running text were available. Second, the methods that will be discussed in the second part of this chapter take advantage of additional information about the informants, namely their gender and age. Speakers where either or both are not available were removed from some analyses. Third, a small number of teenagers and children were removed.

About two thirds of the FRED informants are older men that have not lived outside their region for a prolonged amount of time (Hernández 2006: 6); in other words, they are members of the population group preferred by dialectologists: non-mobile older rural men (NORMs). Unfortunately, speaker age and gender are not available for all informants, with age being unknown for 146 speakers and gender missing for 22. Both are necessary for most analyses covered in the later chapters. In some cases, the age can be approximated from the text, or from additional information such as the combination of birth and interview years or decades. For example, informant A109 from Nottinghamshire was born in 1912, but only the interview decade was recorded (the 1980s), not the exact year in which the interview took place. We can conclude, however, that at the time of the interview A109 was at least 68 and at most 77 years old. I therefore place this speaker in the middle of the range, at 73 years of age. Where this was not possible, the whole text had to be removed from consideration. An exception to this are the bottom-up n-gram analyses. Here, geographic, age-based and gender-based distributions are analyzed separately, and speakers were only removed from those analyses for which the metadata was missing.

Finally, a small number of very young informants were removed. This usually applies to family members, children or teenagers, that were present during the interview and contributed a small amount of spoken material. Including these speakers might lead to wrong estimates for the effect of age, as they are clear outliers. The threshold for inclusion was selected as 40 years younger than the average age, i.e. at an age of about 31.

After these exclusions, 273 informants and about 2.1 million words remain. The next section gives more information on the geographic and sociological characteristics of the data set.

3.1.1.1. Areal coverage

The following section details the 38 counties included in FRED, and to what extent they are covered both in the data sets under study.

Angus Chapman code: ANS; FRED region: Scottish Lowlands
FRED 5 speakers with 19900 words in total. 100 percent male speakers, mean age 78.2
FRED-S No data available

Banffshire Chapman code: BAN; FRED region: Scottish Lowlands
FRED 1 speakers with 5655 words in total. 0 percent male speakers, mean age 76.0
FRED-S No data available

Cornwall Chapman code: CON; FRED region: Southwest of England
FRED 10 speakers with 97766 words in total. 100 percent male speakers, mean age 72.2
FRED-S 6 speakers with 27240 words in total. 83 percent male speakers, mean age 80.0

Denbighshire Chapman code: DEN; FRED region: Wales This county is neither included in FRED not in FRED-S for this study.

Devon Chapman code: DEV; FRED region: Southwest of England
FRED 7 speakers with 61280 words in total. 86 percent male speakers, mean age 83.6
FRED-S 11 speakers with 81532 words in total. 64 percent male speakers, mean age 83.0

Dumfriesshire Chapman code: DFS; FRED region: Scottish Lowlands
FRED 1 speakers with 9997 words in total. 100 percent male speakers, mean age 72.0
FRED-S No data available

Durham Chapman code: DUR; FRED region: North of England
FRED 3 speakers with 28069 words in total. 67 percent male speakers, mean age 78.3
FRED-S 3 speakers with 27008 words in total. 67 percent male speakers, mean age 78.3

East Lothian Chapman code: ELN; FRED region: Scottish Lowlands
FRED No speakers included
FRED-S 11 speakers with 29403 words in total. 36 percent male speakers, mean age 17.8

Glamorganshire Chapman code: GLA; FRED region: Wales
FRED 6 speakers with 47365 words in total. 100 percent male speakers, mean age 81.7
FRED-S No data available

Hebrides Chapman code: HEB; FRED region: Hebrides
FRED 13 speakers with 49574 words in total. 46 percent male speakers, mean age 65.4
FRED-S No data available

Isle of Man Chapman code: MAN; FRED region: Isle of Man
FRED 2 speakers with 10930 words in total. 100 percent male speakers, mean age 81.0
FRED-S No data available

Kent Chapman code: KEN; FRED region: Southeast of England
FRED 9 speakers with 176233 words in total. 100 percent male speakers, mean age 84.9
FRED-S 9 speakers with 157701 words in total. 89 percent male speakers, mean age 85.1

Kincardineshire Chapman code: KCD; FRED region: Scottish Lowlands

FRED 1 speakers with 5733 words in total. 100 percent male speakers, mean age 71.0

FRED-S No data available

Lancashire Chapman code: LAN; FRED region: North of England

FRED 23 speakers with 205326 words in total. 52 percent male speakers, mean age 67.0

FRED-S 13 speakers with 141749 words in total. 38 percent male speakers, mean age 72.4

Leicestershire Chapman code: LEI; FRED region: English Midlands

FRED 1 speakers with 5864 words in total. 100 percent male speakers, mean age 72.0

FRED-S No data available

London Chapman code: LND; FRED region: Southeast of England

FRED 6 speakers with 108977 words in total. 50 percent male speakers, mean age 65.2

FRED-S 8 speakers with 77277 words in total. 38 percent male speakers, mean age 62.5

Middlesex Chapman code: MDX; FRED region: Southeast of England

FRED 2 speakers with 31795 words in total. 100 percent male speakers, mean age 74.5

FRED-S 2 speakers with 31170 words in total. 100 percent male speakers, mean age 74.5

Midlothian Chapman code: MLN; FRED region: Scottish Lowlands

FRED 2 speakers with 15217 words in total. 100 percent male speakers, mean age 62.0

FRED-S 4 speakers with 21358 words in total. 75 percent male speakers, mean age 56.0

Northumberland Chapman code: NBL; FRED region: North of England

FRED 5 speakers with 30647 words in total. 40 percent male speakers, mean age 81.2

FRED-S 5 speakers with 28429 words in total. 20 percent male speakers, mean age 81.5

Nottinghamshire Chapman code: NTT; FRED region: English Midlands

FRED 16 speakers with 150816 words in total. 62 percent male speakers, mean age 80.7

FRED-S 16 speakers with 136857 words in total. 69 percent male speakers, mean age 80.6

Oxfordshire Chapman code: OXF; FRED region: Southwest of England

FRED 3 speakers with 14357 words in total. 100 percent male speakers, mean age 85.3

FRED-S 4 speakers with 14285 words in total. 75 percent male speakers, mean age 85.3

Peebleshire Chapman code: PEE; FRED region: Scottish Lowlands

FRED 2 speakers with 14956 words in total. 50 percent male speakers, mean age 56.0

FRED-S No data available

Perthshire Chapman code: PER; FRED region: Scottish Lowlands

FRED 4 speakers with 17088 words in total. 100 percent male speakers, mean age 82.2

FRED-S No data available

Ross and Cromarty Chapman code: ROC; FRED region: Scottish Highlands

FRED 2 speakers with 10475 words in total. 50 percent male speakers, mean age 80.0

FRED-S No data available

Selkirkshire Chapman code: SEL; FRED region: Scottish Lowlands

FRED 3 speakers with 9325 words in total. 100 percent male speakers, mean age 65.0

FRED-S No data available

Shropshire Chapman code: SAL; FRED region: English Midlands

FRED 31 speakers with 149987 words in total. 77 percent male speakers, mean age 81.3

FRED-S No data available

Somerset Chapman code: SOM; FRED region: Southwest of England

FRED 28 speakers with 176690 words in total. 96 percent male speakers, mean age 80.1

FRED-S 13 speakers with 66922 words in total. 92 percent male speakers, mean age 79.0

Suffolk Chapman code: SFK; FRED region: Southeast of England

FRED 30 speakers with 295339 words in total. 100 percent male speakers, mean age 74.1

FRED-S No data available

Sutherland Chapman code: SUT; FRED region: Scottish Highlands

FRED 4 speakers with 10967 words in total. 100 percent male speakers, mean age 66.5

FRED-S No data available

Warwickshire Chapman code: WAR; FRED region: English Midlands This county is not included in either FRED or FRED-S for this study.

West Lothian Chapman code: WLN; FRED region: Scottish Lowlands

FRED 3 speakers with 16410 words in total. 67 percent male speakers, mean age 66.7

FRED-S 4 speakers with 16520 words in total. 50 percent male speakers, mean age 66.7

Westmorland Chapman code: WES; FRED region: North of England

FRED 20 speakers with 151806 words in total. 60 percent male speakers, mean age 79.0

FRED-S 5 speakers with 21591 words in total. 80 percent male speakers, mean age 78.6

Wiltshire Chapman code: WIL; FRED region: Southwest of England

FRED 21 speakers with 152161 words in total. 57 percent male speakers, mean age 77.0

FRED-S 16 speakers with 76499 words in total. 50 percent male speakers, mean age 74.1

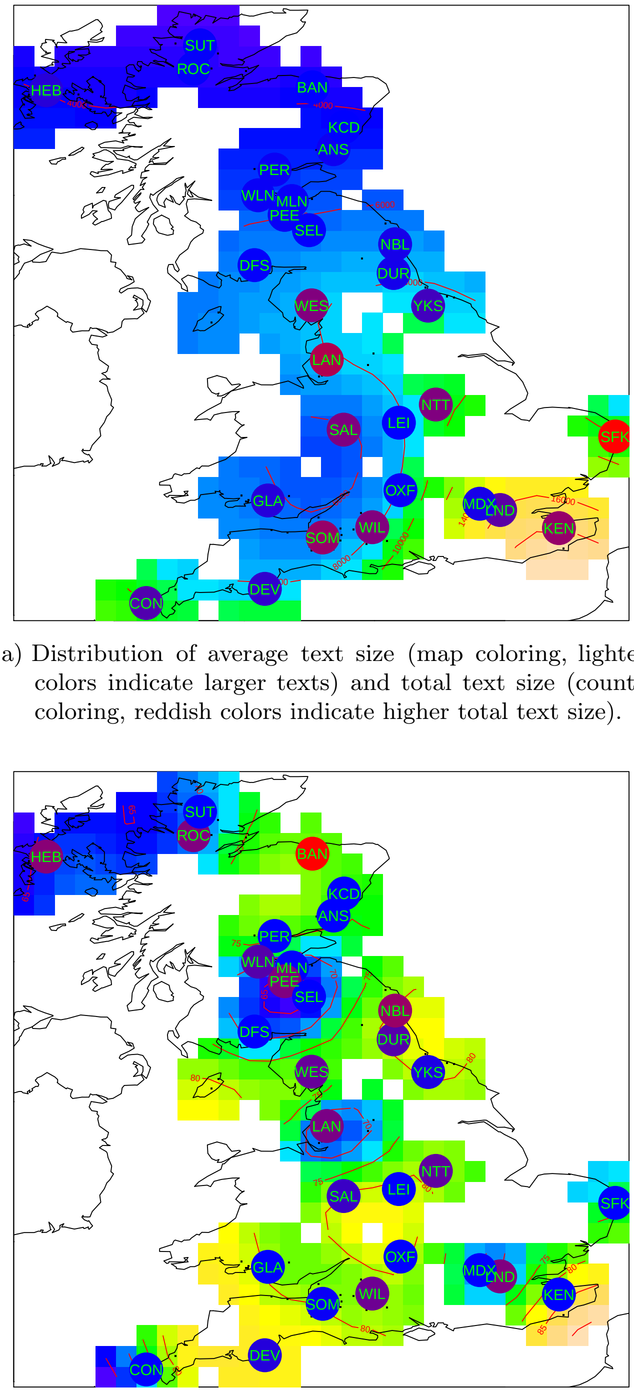
Yorkshire Chapman code: YKS; FRED region: North of England

FRED 9 speakers with 79614 words in total. 89 percent male speakers, mean age 81.6

FRED-S 11 speakers with 52672 words in total. 64 percent male speakers, mean age 78.8

Map 2a visualizes the distribution of the data in FRED. The colors in the background of the map indicate the average number of words per speaker, with blue indicating fewer words. Clearly, most counties have similar averages of around 6,000-8,000 words. The general pattern is such that more northern counties have shorter texts, and the southern counties, especially those in the Southeast, have longer ones. The total number of words per county is indicated by the colors of the county marker. The colors match the position of the county on a scale running from the county with the smallest amount of words (Banffshire) in the deepest blue, to that with the highest amount of words (Suffolk) in red. Intermediate counties exhibit proportional shades of purple. The counties in the English South and Midlands have, with a few exceptions such as Middlesex, relatively good coverage. For Scotland and the northeastern parts of the English North, on the other hand, only little data is available.

Map 2a similarly visualizes the geographic distribution of the sociological factors. The background coloring of the map represents the average speaker age. Generally, this lies at around 75 years. The southern Scottish Lowlands have particularly young speakers, averaging below 70 years; Lancashire and the area around London have relatively young informants as well. Northumberland, Kent, and the Southwest of England have the on average oldest informants. The colors of the county marker indicate the gender distribution there: the more reddish that color is, the greater the proportion of female informants.



(a) Distribution of average text size (map coloring, lighter colors indicate larger texts) and total text size (county coloring, reddish colors indicate higher total text size).

(b) Distribution of speaker age (map coloring, lighter colors indicate older average speaker age) and gender (county coloring, reddish colors indicate more female speakers).

Map 2: Text size, age and gender distribution in FRED.

The North of England and the Midlands generally have more female speakers, while the Southwest has mostly male speakers. Scotland presents a mixed picture.

3.1.2. The feature set

Tapping the FRED corpus, Szmrecsanyi (2013) derived a feature list containing 57 dialectologically relevant features and extracted the relevant feature counts. For the model-based analyses in Section 4.1.1, I re-analyze his data. A description of the features involved, together with information on the extraction process and the distributional results, can be found in that section and in greater detail in Szmrecsanyi (2013: Section 3.4), and, concerning the technicalities, in Szmrecsanyi (2010a).

One crucial difference between Szmrecsanyi's list and the one used for the present study concerns how binary alternations are treated. In the original study, they were always included as two separate frequency vectors, one for each realization. Here, they are combined into a single feature representing the choice between them. The original list contains 12 such alternations, therefore the data now covers 45 different features. Table 3.1 gives the full list.

3.2. Methods

This section discusses the methods used in the present study. It consists of two parts: first, I extend the method proposed by Szmrecsanyi (2013) using two different components: *mixed-effect modeling* and *generalized additive modeling*. Section 3.2.2 will describe these methods and provide a discussion of their validity and feasibility. An expose of Szmrecsanyi's method will precede this in Section 3.2.1, to provide the necessary methodological background. Then, a radically different method will be introduced in Section 3.2.3. Instead of starting from a carefully selected feature list, this method proceeds in a bottom-up fashion and aims to discover and measure interesting features directly from the data itself. The results of both processes will be covered in Chapter 4 and used for dialectometric purposes in Chapter 5.

3.2.1. Corpus-based dialectometry

This section describes the original CBDM methodology, following (Szmrecsanyi 2013: Section 2.2 and 3.2). It can be thought of as a simple step-by-step process, akin to a "cooking recipe" (26). The process starts with a dialect corpus, and as a first step defines a feature catalog. This catalog is constructed by selecting a list of features where the

1/2	(Non-)standard reflexives
3/4	The archaic pronouns *thee, thou, thy*
4	The archaic pronoun *ye*
5	*us*
6	*them*
7	synthetic adjective comparison
8/9	the genitive alternation
10	preposition stranding
11/12	cardinal number + *year(s)*
13	The primary verb *to do*
14	The primary verb *to be*
15	The primary verb *to have*
16	marking of possession: *have got*
17/18	Future markers *be going to* and *will* or *shall*
19/20	habitual past: *would* or *used to*
21	progressive verb forms
22/23	present perfect: auxiliaries *be* and *have*
24	marking of epistemic and deontic modality: *must*
25	marking of epistemic and deontic modality: *have to*
26	marking of epistemic and deontic modality: *got to*
27	*a*-prefixing on *-ing* forms
28	non-standard weak past tense and past participle forms
29	non-standard past tense *done*
30	non-standard past tense *come*
31	the negative suffix *-nae*
32	the negator *ain't*
33	multiple negation
34	contraction in negative contexts
36	*never* as past tense negator
37/38	*wasn't* and *weren't*
39	non-standard verbal *-s*
40/41	*don't* or *doesn't* with 3rd person singular subjects
42	existential/presentational *there is/was* with plural subjects
43	absence of auxiliary *be* in progressive constructions
44	non-standard *was*
45	non-standard *were*
46	*wh*-relativization
47	the relative particle *what*
48	the relative particle *that*
49	*as what* or *than what* in comparative clauses
50	unsplit *for to*
51/52	infinitival or gerundial complementation after *begin, start, continue, hate,* and *love*
53/54	zero or *that* complementation after *think, say*, and *know*
55	lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes/no* questions
56/57	the dative alternation following the verb *give*

Table 3.1.: Feature set used for the model-based analyses.

literature suggests a geographic distribution, or where such a distribution is at least in principle possible. The list can be arbitrarily large; it is only constrained by the feasibility of corpus investigation. Szmrecsanyi (2013: 35ff.), in constructing his list, drew on the relevant dialectological, variationist and corpus-linguistic literature, with a particular focus on features included in comparative studies across several varieties. Excluded were exceedingly rare phenomena with fewer than 100 observations in FRED, as corpus-based methods are unlikely to lead to accurate results here, and features that require a particularly extensive amount of manual annotation, as investigating these features on such a large scale is not feasible.

Once these prerequisites are in place, the analyst proceeds by creating the frequency matrix. This involves determining the feature frequencies per location, using the standard methods available in corpus linguistics. After completion of this step, the researcher then makes the data comparable across sites using normalization, i.e. by dividing the number of attestations per location by the amount of words available at that location, then scaling that number up to a reference size. This reference size is here always ten thousand words, and the normalized frequencies are therefore average frequencies per ten thousand words (*pttw*). Szmrecsanyi then recommends a logarithmic transformation (with base 10) "to de-emphasize large frequency differentials and to alleviate the effect of frequency outliers" (25). As the logarithm is not defined for zero, features that do not appear in a given location are set to -1, i.e. 0.1 observations *pttw*. The individual values for the features are then compiled into the frequency matrix, in which the rows represent the locations and the columns contain the log-transformed normalized frequency counts for each feature.

As an example, consider Feature 5, the first person plural object pronoun *us*. In the London subcorpus of FRED, which has a total size of 108,977 words, this feature appears 67 times. Its normalized frequency is therefore $(67/108977) * 10000) \approx 16.6$ *pttw*. After using the logarithmic transformation, we arrive at a final frequency value of about 1.22.

Repeating this process for all features and locations, one finally arrives at the frequency matrix. The next step then aggregates over all features, to calculate precisely how the locations relate to one another with regard to the combination of all of these features. The process for this was already discussed in Section 2.1.2; I will illustrate it here with a small example. Consider Table 3.2a, a subset of the frequency matrix, which shows the values for London (LND), Nottinghamshire (NTT) and Northumberland (NBL) for the following three features: the already mentioned *us*, the negator *ain't*, and multiple negation (Features 32 and 33). In Northumberland, *ain't* is unattested in the corpus, and the final frequency value is set to -1. We can now calculate the distances between the locations by plugging the values into the distance function. For the pair London and

	[6]	[32]	[33]
LND	1.22	0.10	0.78
NTT	1.30	0.20	0.77
NBL	1.12	-1	0.41

(a) Subset of data

	LND	NTT	NBL
LND			
NTT	0.13		
NBL	1.16	1.27	

(b) Resulting distance matrix

Table 3.2.: Distance matrix calculation.

Nottinghamshire, using the Euclidean distance function, this results in the following:

$$\begin{aligned} \mathrm{d}(LND, NTT) &= \sqrt{(1.22 - 1.30)^2 + (0.1 - 0.2)^2 + (0.78 - 0.77)^2} \\ &= \sqrt{0.0064 + 0.01 + 0.0001} \approx 0.13 \end{aligned}$$

Applying this process to all pairwise combinations yields the distance matrix in Table 3.2b. This matrix can then be used as the input for a wide variety of analysis techniques, such as hierarchical clustering algorithms, NeighborNet, or multidimensional scaling, as discussed in Section 2.1.2.

3.2.2. Extending corpus-based dialectometry

Consider again Feature 5, the personal pronoun *us*. In Banffshire, the county in FRED with the smallest amount of text in running words, there are three tokens in slightly more than 5,000 words. In Angus, one of its closest neighbors, there are 19 observations in 19,000 words. In Lancashire, finally, a county with an excellent textual coverage of over two hundred thousand words, there are 420 instances of *us*. Let us use the CBDM methodology on this sample. The normalized value for Banffshire works out to $3/5655 * 10000 \approx 5.3$, for Angus it is $19/19900 * 10,000 \approx 9.5$ and for Lancashire $420/205326 \approx 20.5$. After logarithmic transformation, we end up with the following values: 0.72 for Banffshire, 0.98 for Angus and 1.31 for Lancashire. In other words, as input to the following analysis steps, Angus is considered to be roughly equidistant from both Angus and Lancashire.

Are these distances warranted? The CBDM methodology aggregates over all speakers at one location. Let us thus have a look at within-county variability. Figure 3.1 displays the normalized frequencies for each individual speaker, ignoring those where fewer than 1,500 words are available[1]. The by-county normalized frequency for Banffshire is highlighted in the plot by a grey line. Concerning Angus, three of the five speakers exhibit normalized frequencies that are virtually indistinguishable from the county mean for Banffshire,

[1] This removes some speakers from Lancashire that have no attestations for *us*, and would skew the overall picture.

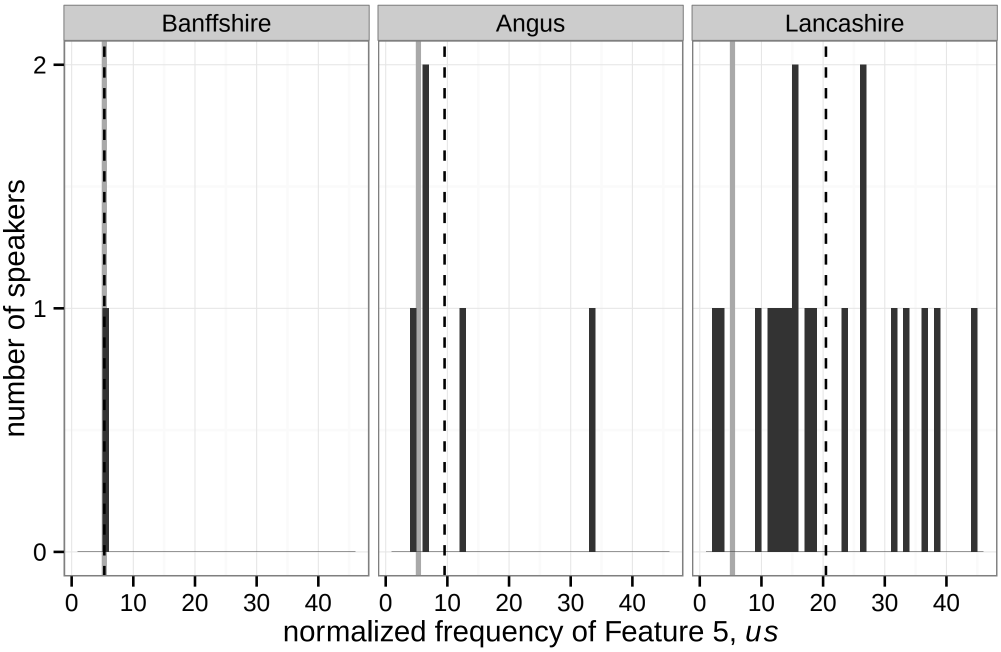


Figure 3.1.: By-speaker normalized frequencies for *us* in Angus, Banffshire and Lancashire. Dashed lines indicate overall normalized frequency per county. Dark grey line shows normalized frequency for Banffshire.

ranging between 4.22 and 6.8. Only two speakers show a much higher rate of usage for this feature. Furthermore, these two have a relatively small number of running words. In Lancashire, on the other hand, most speakers clearly make use of this feature much more frequently. Nevertheless, two of the 19 speakers are clearly below the normalized frequency for Banffshire.

Aggregating over all speakers in a county boils this variability down to a neat number, and in general should make the judgment about how prevalent a certain feature is in a region more accurate. However, this is crucially influenced by how much data we have. If, by some accident in the sampling process of FRED, the two speakers from Angus that show particularly high rates of *us* had been excluded from the corpus, we would conclude that there is very little difference between these counties. The normalized frequency for Angus would be 6.4, or 0.8 after application of the logarithmic transformation, which is quite close to the 0.72 of Banffshire. Thinking about this in another way, how surprised would we be if the speaker from Banffshire actually turned out to be from Angus, as far as this feature is concerned? Three of five speakers there show a very similar rate, so the speaker from Banffshire would fit in rather neatly. For Lancashire, this is different: while a small number of speakers use this feature roughly as often as the informant from Banffshire does, most use it much more often. The spread of frequencies across speakers there does show, however, that even in counties that clearly behave differently, individual speakers are quite variable. Assuming this is also true for Banffshire, it is quite possible that this speaker falls on the lower end of the spectrum, and we again end up with distances that are over-inflated.

Of course, the purpose of CBDM is not to show, in a statistically reliable way, that individual counties are different from one another with regard to a single feature. Rather, it seeks to aggregate the individual measures into a combined value, which smooths over the differences for individual features. Yet, if this influence of relative sample size is a problem, it should be so for many features, and aggregation may well increase the problem instead of reducing it. When the individual measurements are likely to be rather inaccurate, it is not necessary that the observed difference is as often smaller as it is larger.

Another potential problem with using normalized values pertains to the influence of factors other than geography. For *us*, there is a clear gender difference: male speakers, when counted across the whole corpus, use this feature at a rate of about 11.3 *pttw*, while female speakers exhibit about twice that frequency (22.3 *pttw*). This difference is highly significant according to a simple χ^2 test($\chi^2 = 300, p < 0.001$). In other words, we would in general expect to find more observations of *us* when more of a county's informants are female. With regard to *us*, this would confirm the measurement that Banffshire and Angus are quite different, assuming the informant from Banffshire is female whereas all from Angus are male. But this is ultimately accidental: if the situations were reversed and Angus had more female speakers than Banffshire does, how would that influence our judgment of how similar they are? We would need to conclude that the difference is even less well-founded. This aspect is especially important for non-standard morphosyntactic features, as female speakers in general tend to use fewer non-standard forms (Chambers 2003).

One solution to the first problem is to scale the frequency counts by how well-supported they are in the data. In other words (and excluding gender for now), the distances between Angus and Banffshire should be smaller than they appear based on normalized feature frequencies, as the evidence that they actually are different is relatively weak. They should still be considered different, though: after all, the data suggests that they are. Only the size of the difference should be reduced, to reflect that, were we to sample different informants, it is quite likely that we would see a less extreme picture.

There are several ways in which one could do this. First, one could look at all the data for one feature and see how it is distributed. One would then assume that overall, the locations are somewhat similar to each other, and evaluate the data in that light. If one group is different from the other groups, and this is based on relatively little data, one should conclude that this group is most likely dissimilar to the other groups, but that the true difference in absolute numbers is probably lower than what it appears to be on the surface. Strong claims need strong evidence, but for weaker claims the criteria can be relaxed.

The advantage of this approach is that, apart from the assignment of speakers to groups (i.e. counties), it is agnostic about geography. Unless explicitly specified, such a technique would not know anything about how the locations relate to one another, and therefore remain completely neutral. The only thing that would matter is how the individual observations in their group behave and what the rest of the data looks like.

This advantage, however, is also a disadvantage. Is it the case that individual locations should primarily be seen in light of how all data behave, ignoring how they arrange spatially? This seems implausible. When considering the behavior of Scottish dialects, other Scottish dialects are much more likely to be a good base of comparison than, say, the Southeast of England. The effects of this assumption can influence the analysis negatively in at least two ways. Consider, first, the case where we have two neighboring locations that behave similarly with regard to a certain feature, but one is based on a larger amount of evidence. Taking a perspective that is agnostic about geography, one would scale the case with weaker evidence more strongly toward the overall behavior; in other words, the distance between the two neighbors should become larger. And in a certain light, this makes sense: after all, it is much clearer from the data that the first location is different in a particular way than it is for the second case. But this seems contrary to, at least, a dialectologist's intuition. Second, if two neighboring locations both have weak evidence in the same direction, they will be both scaled more strongly toward the overall behavior. Again, this is logical, but contrary to intuition; instead it seems more plausible to assume that close regions, in some way, should count more toward each other than far-away locations do. And precisely this is what the second approach entails. Each location is seen in the light of how it relates to points that are close, and to points that are far away.

Both approaches can be implemented using variants of *(generalized) linear regression modeling*. Such models have the advantage that they can also address the effect of other factors, such as speaker age and gender. For the first approach, I rely on *generalized linear mixed modeling* and the so-called *partial pooling* effect. Here, geography will be included as a simple categorical factor. For the second approach, I turn to *generalized additive modeling*, in which geography will be included directly using smooth functions. Both will be discussed in greater detail in the two following sections.

Regardless of the specific model, the same basic steps are involved. I mostly follow the recipe described in Szmrecsanyi (2013) as outlined in the previous section, but replace the normalization step with one in which a probabilistic model is fit to each feature. These models are then used to predict, for each county, the number of instances that we would expect an idealized average male speaker of the overall mean age to produce in ten thousand words. These values correspond to the normalized frequencies in the original method and are on the same scale. Analysis can then proceed in the same way, i.e. logarithmic transformation and conversion to a distance matrix using the Euclidean distance metric.

The feature models always include terms for the sociolinguistic effects, speaker age and gender, as well as their interaction. Gender was coded as a binary variable, with the most frequent value (male) as the default. Speaker age was centered around the mean. In other words, the base model predictions are for a male speaker of average age. Non-significant terms were not removed from the analysis. This is not customary in linguistic models, which usually strive to find the simplest model that can account for the data. Here, model simplicity is not the main goal, but predictive accuracy. Therefore, I follow the recommendations by Gelman & Hill (2007: 69) for removing terms from such models: "If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them." As for individual features both gender and age may have an effect in any direction, no predictors are removed.

Generalized linear models require an explicit probability distribution that the data are assumed to follow. The choice of this distribution crucially depends on the type of data that the analyst seeks to model. One of the most commonly used distributions is the *binomial* distribution. It is used to model binary outcomes, and the resulting regression models are known as *logistic regression* models. As many linguistic phenomena can be considered binary alternations, this type of analysis has found a large following in linguistic circles. It first emerged among sociolinguists, where a variant named VARBRUL (variable rules analysis) is widely popular to this day (Sankoff & Labov 1979, Tagliamonte 2012). In recent years, it has made inroads in corpus- and psycholinguistic research, especially that with a focus on syntax (e.g. Bresnan et al. 2007, Szmrecsanyi 2010b, Grafmiller forthcoming). Therefore, where features could conceivably be modeled as alternations, I chose to do so. This is in contrast to the analyses presented in Szmrecsanyi (2013), who only considered absolute feature frequencies. The probability mass function for the binomial distribution can be found in Equation (3.1), where k represents the number of times one particular outcome happens, p the probability for that outcome, and n the total number of tries. The logistic regression estimates this probability, and the influence that independent factors have on it, from the data.

$$f(k;n,p) = \Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

This leaves us with the remaining features, which require count-based models. The computational linguistics literature has covered the question of appropriate probability models extensively. The most basic and widely used distribution for count data across most scientific disciplines is the Poisson distribution. Equation (3.2) gives its probability mass function; k here is the number of times a certain outcome happens in a time unit and λ is the parameter to be estimated, indicating how frequent that outcome is. By adding the number of words as an offset[2] to the corresponding model, it can be used to predict rates of occurrence instead of absolute counts. I will show in the next section that normalization is, in essence, a very simple Poisson regression. There is, however, evidence that words in general do not follow this distribution particularly well, as occurrences tend to be more grouped and "bursty" than the Poisson distribution allows (Altmann et al. 2009, Pierrehumbert 2012). It has already been established in the 1970s, though, that more grammatical items approximately follow this distribution (Bookstein & Kraft 1977), and Manning & Schütze (1999: 547) consider it "good for non-content words". Altmann et al. (2009) recommend the Weibull distribution, where probability decreases by an additional parameter indicating distance from the last occurrence, but this method is difficult to implement with the present tools and data set. Manning & Schütze (1999) also suggest the negative binomial distribution as a more adequate choice. Its probability mass function can be found in Equation (3.3). It has one additional parameter compared to the Poisson distribution, which allows the model to account for more variation in how the data are dispersed.

$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.2}$$

$$f(k;r,p) \equiv \Pr(X = k) = \binom{k+r-1}{k} (1-p)^r p^k \quad \text{for } k = 0, 1, 2, \ldots \tag{3.3}$$

The choice of the probability distribution in the present study is constrained by what is available in the statistical packages used here. The binomial distribution is available for both model types. Concerning the other distributions discussed above, lmer, the method

[2] An offset is a predictor in the model that does not have its effect estimated, but is fixed to one.

implementing the first approach, offers the Poisson distribution as the only choice. For the second approach, the GAM package `mgcv` allows the Poisson as well as the negative binomial distribution. I chose to use the negative binomial distribution for these models to handle potential overdispersion at least in one set of results.

3.2.2.1. Leveraging the partial pooling effect using lmer

Generalized linear mixed modeling Pinheiro & Bates (2000), also referred to as multilevel modeling, is an extension of generalized linear modeling. Gelman & Hill (2007: 1) give a concise definition:

> [W]e consider a multilevel model to be a regression (a linear or generalized linear model) in which the parameters–the regression coefficients–are given a probability model. This second-level model has parameters of its own–the hyperparameters of the model–which are also estimated from data. [. . .] The feature that distinguishes multilevel models from classical regression is in the modeling of the variation between groups.

To understand the effect of this on the results, it is instructive to consider three concepts that Gelman & Hill (2007) introduce: *complete pooling*, *no pooling*, and *partial pooling*. Complete pooling refers to analyses that ignore between-group differences completely, and therefore can only give the overall mean as a result. The estimation of this general trend, however, is very accurate, since all of the data can be taken into account. Compare this to no pooling, where a separate model is fit for each group; for present purposes this is equivalent to including the term as a regular (i.e. fixed effect) predictor in the model. In that case, the estimations for each group are made based only on the observations for that group, and therefore resemble the available data as closely as possible. Information about the other groups and their variability, however, is discarded.

Partial pooling is a hybrid of these two approaches. The influence that the grouping has on the data is included in the analysis, but the effect size of the grouping is constricted by a probability distribution. For the mixed-effect models used here, this distribution is the simple normal distribution. In essence, this means that the groups are assumed to be drawn at random from a larger population, and that the distribution of the variability among that population follows a Gaussian bell curve. The parameters of this curve are themselves drawn from the distribution of the data. The influence of each group is then predicted based on the data and this probability distribution. This leads to what is known as *shrinkage* (cf. Baayen 2008: 274ff.): compared to the no pooling estimates, each value is shrunk toward the overall mean, proportional to the amount of data that is available for that value. Thus, such models "yield more reliable estimates of group-specific properties" (Jaeger et al. 2011: 313). The per-group values are referred to as BLUPs (Best Linear Unbiased Predictors).

Mixed-effect regression models can include both random effects, i.e. predictors using partial pooling, and fixed effects that correspond to no pooling. As the fixed effects are similar to those in a classical regression model, they can be evaluated for their statistical significance. This characteristic has found many applications in linguistic research, where factors abound that may influence the analysis, but by themselves are not crucial to the goal of the study. For example, in a psycholinguistic experiment on whether a particular part-of-speech ambiguity of lexical items has an effect on processing time, there are a lot of external factors that could influence the results. To name just two, individual participants may be faster than others, or individual items may be more difficult than others. While possibly influential, these effects are itself not important to the study. Ignoring such variation may make inferences invalid (Clark 1973), but including them as a fixed effect has many problems, such as greatly inflating the complexity of the model and making predictions to new data very hard (cf. Baayen 2008: 241). Mixed-effect modeling has emerged as one of the best solutions to this problem (Baayen et al. 2008, Quené & van den Bergh 2008, Jaeger 2008). Similarly, in grammar-oriented research they have been used to account for potential differences between individual speakers, lexemes or registers (Bresnan & Ford 2010, Wolk et al. 2013). Sociolinguists have also begun to use them to account for the fact that individual speakers may exhibit idiosyncratic variation (Tagliamonte & Baayen 2012, Johnson 2009).

The CBDM approach using normalization is equivalent to a no pooling estimate. This can be seen as follows: the normalization process uses Equation (3.4), where x_i refers to the number of occurrences of a given feature by speaker i in county x, and X_i refers to the number of total observations (i.e. words). The maximum likelihood estimator for Poisson-distributed data is given in Equation (3.5), where y_i represents the number of events for one time unit, and N the number of observations. As texts have different numbers of words, we cannot use them as the time unit. Instead, let us use words as the unit of time, and identify each instance of a feature use with exactly one word, so that y_i equals 1 when word i is an instance of that feature, and 0 if not. Next, we group words by speakers in both numerator and denominator, and we end up with exactly Equation (3.4), save for the normalization constant.

$$\text{norm}(x) = \frac{x_1 + x_2 + \ldots + x_i}{X_1 + X_2 + \ldots + X_i} * 10000 \tag{3.4}$$

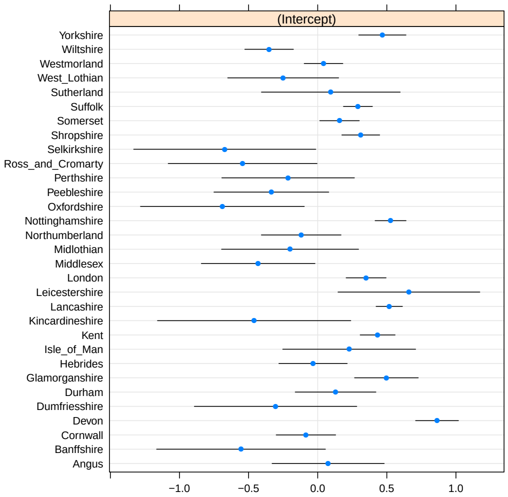


Figure 3.2.: BLUPS (intercept adjustments) by county for Feature 5: *us*

(Intercept)	-7.043^{***}
	(0.094)
Sex: female	0.629^{***}
	(0.053)
Age (centered)	-0.000
	(0.003)
Sex × Age	0.011^{**}
	(0.004)
county (random effect)	0.213
	(0.461)
N	273
Groups	31

Table 3.3.: lmer model for Feature 5: *us*. For fixed effects, coefficients and standard errors (in parentheses) are displayed. Positive coefficients indicate higher frequency.

$$\hat{\lambda} = \frac{y_1 + y_2 + \ldots + y_i}{N} \tag{3.5}$$

Thus, when performing the normalization step, the analyst is in essence doing a basic Poisson regression. It is, however, using the no pooling approach to variation between counties, and in contrast to explicit modeling cannot easily include other predictors. Therefore, the model was made explicit and implemented, for each feature, using the software package `lmer`. All social predictors were included as fixed effects, and county was modeled as a random effect.

The result of such modeling for *us* is shown in Table 3.3. The row labeled *(Intercept)* gives the natural logarithm of the expected overall rate, −7.043, i.e. for a default speaker we would expect a rate for *us* of $e^{-7.034}$, which works out to 8.81 observations *pttw*. Unsurprisingly, this value is significantly different from 0, as indicated by the standard error and the p value in the line below. This means that we can reject the hypothesis that every word is an instance of *us* with very high probability. The next value introduces the first sociolinguistic effect. Female speakers add 0.629 to this intercept; this means that overall we would expect women to use *us* $e^{0.629} = 1.8$ times as often as men do, or 16.5 *pttw*. Again, this value is highly significant, which shows that this frequency difference seems reliable. This is in contrast to the coefficient for age, which is both very small and not significant. In other words, for male speakers, age does not affect the frequency of this feature. The interaction term of both sociolinguistic predictors is significant again; this means that older female speakers do behave differently from younger female speakers. The coefficient is positive, which means that older women use *us* more often than younger women do. Each year adds the sum of both age terms to the resulting value, so female speakers that are ten years older than the average age would produce $e^{(-0.000+0.011)*10} = 1.12$ times the number tokens than the female speaker of average age.

Figure 3.1 shows the distribution of BLUPs. Table 3.3 states that this factor has a standard deviation of 0.46, and the values fit the normal distribution quite well, according to a Shapiro-Wilk test ($p > 0.6$). The positions can be treated like the factors above; for example, Angus is very close to the center of the distribution and shows only slightly higher rates for this feature than average: $e^{0.07}$ times the normal amount, i.e. 7 percent more. In contrast, the feature is used much more often in Devon, where the model predicts $e^{0.863}$, i.e. 2.37 times the average amount. Oxfordshire on the other hand only exhibits this feature very rarely, with a rate of $e^{-0.689}$, i.e. about half of the overall average.

This, of course, raises a concern: are these values "better" than the ones yielded by simple normalization, and can we prove this? Furthermore, the models assume that the between-group variability is normally distributed. While for many features this seems to be true, as we have seen for Feature 5 above, in other cases it is obviously not. A typical example would be Feature 31, the negating suffix -*nae*, which is almost completely restricted to Scotland yet quite frequent there. Therefore, the frequency for most counties will be zero, but in a few it will be rather high. It follows that the variability cannot be normally distributed, which violates the assumptions of the model. It will need to be tested whether this has adverse effects on the process. Johnson (2009: 380) notes that

> [n]ormality of random effects is also an assumption of mixed-model analysis. In practice, the mixed model does not require its random effects to be normally distributed. If they are not, however, the quality of inference that can be made from the model suffers.

In principle, it is not possible to know whether the lmer BLUPs are better for any specific feature than the normalized values are. To do so would require knowing how the dialects "actually are" with regard to this feature, and if that was a known quantity, then no modeling would be necessary: one could simply use the true values as the input for further analysis. We can, however, test how likely it is that using this method leads to better results, using a simulation-based approach.

Let us consider that we had a feature where we knew the true values. We could then compare the two methods by calculating the correlation coefficient between the two. As we are primarily interested in the effect on distance measurements, we would correlate the distances between counties. The better a value is correlated with the true values, the more favorable we would judge that method. Unfortunately, we do not know this for our real data. On the other hand, finding *possible* true values for features is not that difficult: in principle, almost everything is possible (even if some things may be unlikely), and therefore we can just choose them at random. Now we have the "true values", but not the data that result from these values. Both methods assume that the individual observations are approximately Poisson-distributed. Therefore, we can create the corresponding data ourselves, by drawing them from a Poisson distribution with the true value for that location as the parameter λ. This yields a data set where we know what the best result should be, and we can therefore proceed as above.

The advantage of this method is that the effect of several factors on the result can be assessed. In particular, the following conditions will be tested:

base frequency How frequent is the feature overall? This condition has the following values: 0.5, 1, 5, 10 and 15 observations *pttw*, a range covering almost all features in this study.

group variability How much do the groups differ on average? This condition has the following values, in standard deviations: 0.1, 0.2, 0.5 and 1. Again, this covers most of the variability in the actual data.

aggregation One of the fundamental observations of aggregational analyses is that using multiple features at the same time smooths over noise and therefore leads to better results. To test this, several features were created with identical distributions. The conditions were one, three, and five different features with the same distribution.

non-normality: difference What if the variability is not normally distributed? In Britain, a bi-modal distribution seems often quite probable, as both the previous literature and the results reported by Szmrecsanyi (2013) note that much of the variability is structured around the kernels England and Scotland. To test the effect of this, some of the locations were drawn from another distribution that differed by a certain amount. This amount is either zero, indicating two equal distributions, three or five times the frequency of the lower frequency group.

non-normality: distribution This factor is an addition to the previous one and is concerned with the number of locations in the high-frequency group. A location was assigned to this group randomly, with a probability of either 0.1, 0.25 or 0.5.

For each combination of parameters, the simulation was run 25 times, yielding 13,500 observations. In each run, a new pseudo-corpus structure was created at random, based on the group size and text size means and standard deviations from FRED. Figure 3.3 displays the results graphically. To reduce the visual complexity, some combinations of conditions were excluded: group variabilities of 1 standard deviation, non-normal differences of more than three, and non-normal group effect sizes of 10 percent, and aggregations over three features with the same information. The plots for these are all essentially similar to the closest ones shown. Furthermore, the non-normality distribution condition is not interesting in the case of equal distributions as both groups are the same, and therefore only one condition is included. In each cell of the plot, the x-axis represents the feature frequency per 10.000 words, and the y-axis shows the correlation between the true distances and those based on normalization or `lmer` modeling. A linear smoother is included to highlight the general pattern. The horizontal distribution of cells displays the effect of parameter settings for between-group variability, and the vertical distribution represents the effect of the various parameters for non-normality and for aggregation. Clearly, both methods show improved results across the board as the base frequency increases, as indicated by the positive slope of the lines. Furthermore, across conditions, the lmer model seems to achieve a better fit; this holds in 11,558 runs, or 86 percent. Regarding the other predictors, increases in group variability also have a notable effect: overall, the lines in the second and third column are higher than those in the first. Increases in the difference between the modes (the second row compared to the third row, and the fifth compared to the sixth) also lead to improved accuracy, as does an increase in the number of features (rows 4–6 compared to 1–3). These results can be tested using regression models. Table 3.4 shows the results for linear regression on the correlations for both lmer and normalization results. For each model, terms that were neither significant by themselves nor in interaction with base frequency were removed. The first two columns show the results of linear regressions for the correlation values for lmer model and normalization values. As the positive coefficients indicate, increases in all predictors improve accuracy for both. For the interactions with base frequency, the sign is negative, indicating that the improvement grows smaller as base frequency grows larger. The third column shows a logistic regression predicting whether the lmer models fare better than simple normalization. A similar story holds: all significant parameters improve the lmer model predictions compared to the normalized values, and this effect decreases with increasing base frequency.

Factor	lmer	normalized	lmer better
Intercept	0.41	0.28	0.92
log base frequency *pttw*	0.16	0.17	0.56
group variability	0.25	0.28	0.31
difference between modes	0.18	0.22	0.00 n.s.
proportion of second mode	n.s.	n.s.	2.07
number of features	0.01	0.02	n.s.
base frequency:group variability	-0.06	-0.05	-0.28
base frequency:mode difference	-0.05	-0.05	-0.10
base frequency:mode proportions	n.s.	n.s.	-0.50
base frequency:features	-0.00	- 0.00	n.s.

Table 3.4.: Simulation results: effects of parameters on measure accuracy. First column: linear regression model predicting match between lmer model distances and true values. Second column: linear regression model predicting match between normalized distances and true values. Third column: logistic regression model predicting whether the lmer model fares better than the normalized values. Positive values indicate a better fit, or (third column) better odds for the lmer model. All coefficients significant at $p < 0.001$ unless marked.

In short then, it is not guaranteed that the lmer modeling process leads to improved results for individual features. It is, however, quite likely to do so, and this is most crucial for features that are infrequent or where the variability between groups is real, but small. Non-normal between-group variability, or at least bi-modal between-group variability, also does not lead to worse results for the lmer model, despite the fact that the assumptions of the model are violated. The analysis can therefore proceed as planned.

3.2.2.2. Representing geography with generalized additive modeling

The method presented in the last section, mixed-effects modeling with county as a random effect, does not include any information about how the counties relate to one another spatially. Including such information in a regular or mixed-effects linear model is in principle possible. For example, the analyst could include the longitude and latitude of the counties as a predictor. The problem with this approach is that it requires the analyst to specify the functional form of the geographic effect. If both terms are included directly, the model can only evaluate a linear gradient along the north-south and east-west axes. Geographic language variation, however, is not constrained in such a way, and putting an *a priori* shape to this variability is undesirable.

A modeling strategy that is more germane to the particular characteristics of geolinguis-

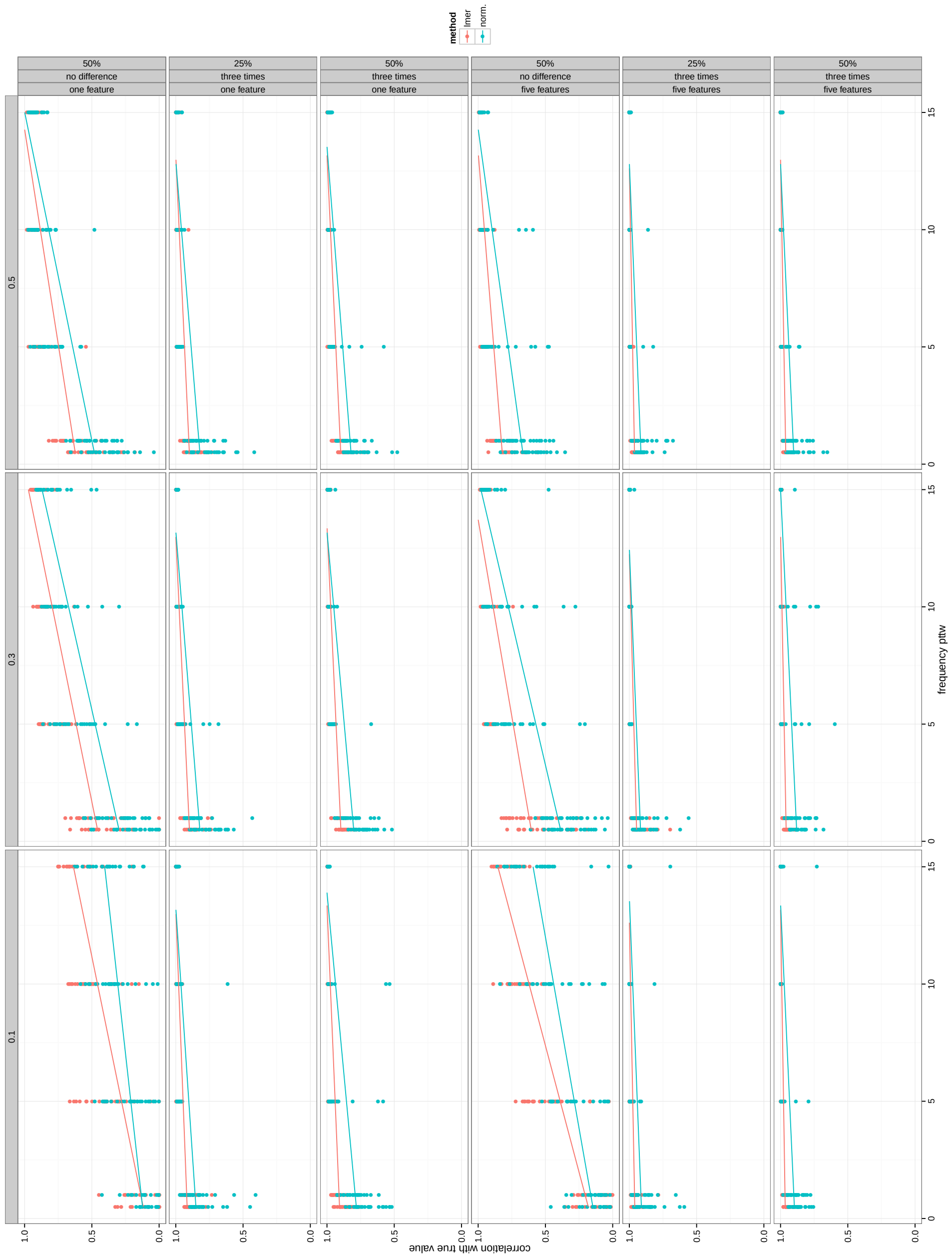


Figure 3.3.: Visualization of simulation results, plotting the base frequency of the feature (x-axis) against the correlation between results and true values (y-axis). Normalization-based results are shown in blue, lmer-based results in red. Columns display the effect of increases in group variability. Rows 1–3 involve a single feature, rows 4–6 five equivalent features. Rows 1 and 4 show a normally distributed variance, rows 2–3 and 5–6 show the effect of a bi-modal distribution (3 and 6: equal size, 2 and 5: 3:1 split).

tics is *generalized additive modeling* (Wood 2006). It seeks to represent complex patterns as a sum of mathematically well-behaved smooth functions. The shape that emerges from this depends, if the functions are chosen correctly, only on the data. *Thin plate regression splines* (Wood 2003) are a good choice for geographic applications, including dialectology (Wieling 2012: 88). The GAM process as applied here makes use of a variation of GCV, a form of *cross-validation*, where measurement points are left out of the analysis, and the model is re-fit and applied to the excluded data. Then, an error analysis weighs the accuracy of the model fit of the remaining and excluded points. This way, the result is less likely to overfit the data and more likely to uncover the true signal.

For the GAM-based analyses in the present work, thin-plate regression splines were used, and GCV was the method for evaluating smoothers. GCV can occasionally lead to extreme values in some areas of the plot; an example of this can be seen in Map 9 (page 89) for the Hebrides. Other options, such as restricted maximum likelihood, were tested and led to less extreme fits in such cases. Overall they lead to oversmoothing, i.e. they had too little flexibility and abstracted away from the data to such a degree that variation between locations mattered too little. The extreme values are not a problem for the aggregational component, as the they are all negative, i.e. close to zero, and the CBDM method enforces a minimum frequency of 0.1 observations *pttw*. For these models, the interviews were not aggregated on the county level. Instead, each interview location was represented using the actual coordinates; this allows the model to be attentive to geographic effects even within counties. To create per-county values for the aggregational step, the predictions were made according to the mean county coordinates.

Figure 3.4 shows the result of this modeling process for Feature 5, *us*. On the x and y axes, we see longitude and latitude, while the z-axis gives the frequency adjustments. In this case, we see a picture emerge that is quite like a mountain range, with high peaks and deep valleys, but also areas of similar frequencies that bundle together, such as the blue areas in the northern part. Like actual mountain ranges, we can plot these frequency mountains in two dimensions as a topographic map. Map 3 shows the result. We can now see clearly that the Scottish Lowlands form a relatively homogeneous region, and that there is a relatively steep frequency boundary running through the North of England. The Midlands again form an area of similar frequencies, while the South exhibits a very complex pattern, with Devon and the Southeast showing much higher frequencies and a steep frequency boundary forming around Wiltshire, Oxfordshire, and Middlesex.

What is the dialectological explanation of this? Let us turn to the *Linguistic Atlas of England* (Orton et al. 1978: M75), who found usage of *us* as a possessive determiner ("with us eyes") confined to a region in the Midlands, which may account for the area of high frequency of *us* there. The field-workers' notebooks for the *Survey of English Dialects* show a particular density of attestations for *us* as a subject pronoun in Devon, intermediate numbers for Cornwall, and low numbers for Somerset and, especially Wiltshire (cf. Wagner 2002: Table 1.2). This matches the pattern in the present map. Kortmann & Wagner (2010) provide a summary map of the distribution of pronoun exchange in the materials in Ellis (1889); for *us* replacing *we*, this is largely restricted to the area around Devon and parts of the Midlands. There is also a certain similarity to Map 185 in (Upton et al. 1987), comparing the distribution of *we two* as opposed to *us two*; the low-frequency areas in Map 3 correspond quite well to those where *we two* predominates. Finally, Wales (2006: 186) notes that *us* for *me* is a widespread feature in the English Northeast and *us* for *our* appears in Yorkshire.

In short, the geographic distribution as estimated by the GAM makes sense linguistically. And importantly, while this method smooths over frequency differences between close points, when a difference is well-supported this will still show up as relatively sharp boundaries. This can be seen, for example, in the transition around London and Middlesex for *us* – one of the features that contributed most to the outlier status of Middlesex in Szmrecsanyi (2013: 133). This leads to representations with straightforward dialectological interpretations, where level sections represent stable areas and sharp increases or decreases represent frequency boundaries. Therefore, GAMs seem very suited for dialectological analysis. Pioneering works in this field are Wieling et al. (2011) and Wieling (2012).

3.2.3. Automated bottom-up syntactic classification

A crucial component of dialectometric analysis is the development and application of measures of linguistic diversity, considering both single features and their aggregated whole. CBDM is an example of this. Automated measurement that proceeds with as little intervention by the analyst as possible can be especially enlightening. Doing so provides the strongest contrast to expert judgments, or at least forces assumptions to be explicitly stated. For measuring phonological and lexical differences, advanced methodologies exist that have proven successful on many data sets; a brief introduction can be found in Section 2.1.1. For the profiling of morphosyntactic variation, however, automated measurement is still in its infancy. Dialect corpora seem to be the most promising data source for this type of analysis. In this section, I present a method for this that is founded on the permutation-based method developed by Nerbonne & Wiersma (2006) as it was applied to dialectometry by Sanders (2007; 2010).

The central issue in automatically measuring syntactic differences is the operationalization of the syntactic dissimilarity inherent in a particular data set. Raw naturalistic

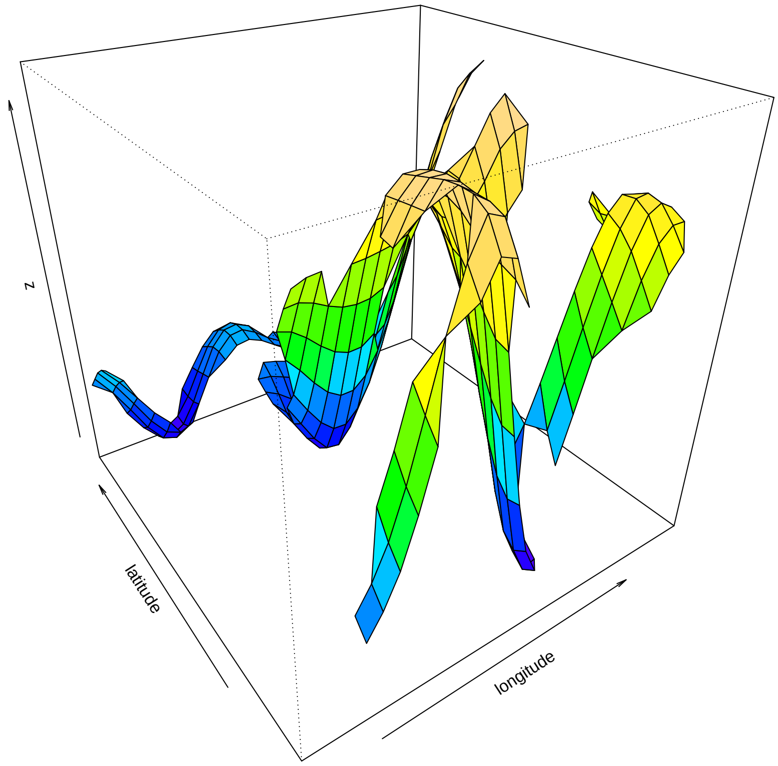
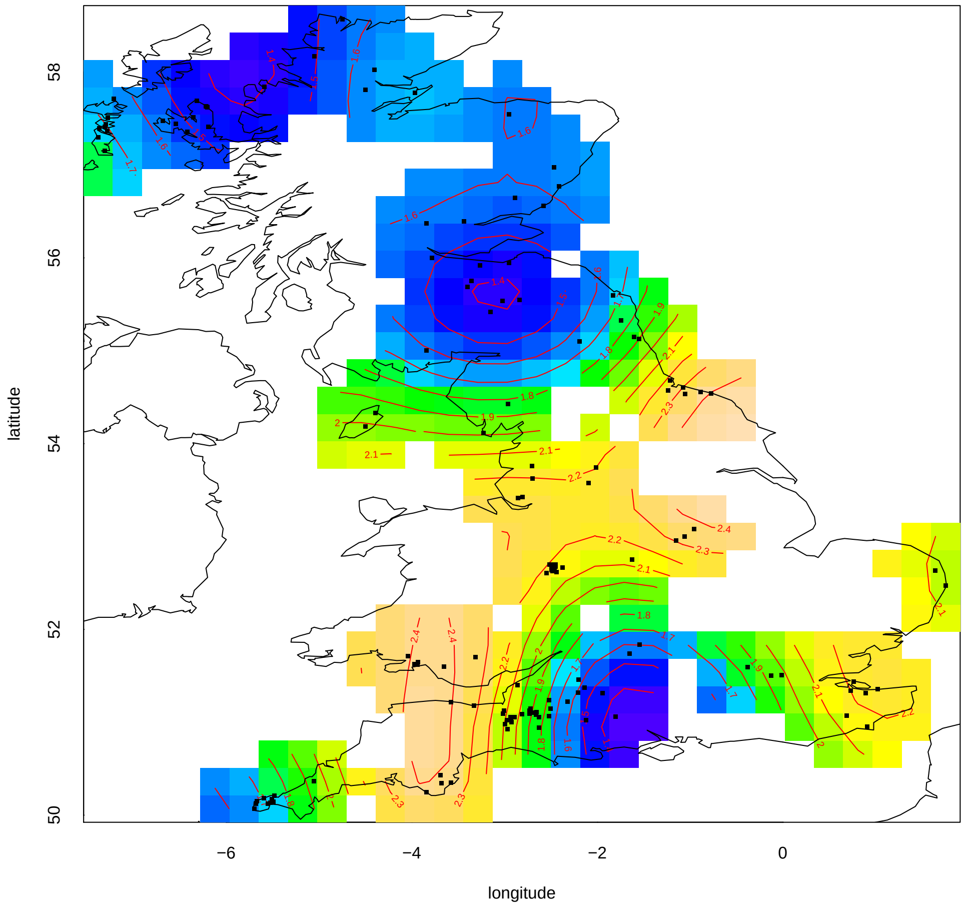


Figure 3.4.: GAM perspective plot for [5]: *us*. Frequency adjustments (z-axis) plotted against longitude (x-axis) and latitude (y-axis). Yellow colors indicate higher frequency, blue colors lower frequency.



Map 3: Geographic effect in the generalized additive model for Feature 5: *us*. Yellow colors indicate higher frequency, blue colors lower frequency.

corpus material is generally unsuitable for this task, as the surface form is strongly influenced by lexical variation and other incidental differences, such as the topics covered in individual interviews. It follows that the analysis needs to proceed on a higher level of syntactic abstraction. While the level of detail for such abstractions may greatly vary, syntactic corpus annotation usually ranges from annotation via part-of-speech (POS) tags to complex syntactic trees according to various formal grammars. POS tags clearly have the least amount of syntactic detail, but this leads to high precision; for example, the `CLAWS4` tagger[3] that annotated the FRED-S corpus generally has an accuracy of 96–97 percent (Garside & Smith 1997: 120). One way of approximating local syntactic contexts in a POS-tagged corpus is the construction of POS n-grams, i.e. all linear POS sequences of a certain length *n*. For linguistic analysis, n usually ranges from 1 (unigrams) to 3 (trigrams), as larger values of n result in greatly increased numbers of n-gram types and thus sparse results. As an example, consider sentence (3) for $n = 2$ (i.e. bigrams). The sentence consists of the second person plural pronoun (with the POS tag `PPIS2`), a past tense lexical verb form (`VVD`), a preposition (`II`), a cardinal number (`MC`), an interjection (`UH`) and punctuation. For this study, punctuation is ignored to avoid effects resulting from transcription differences. Combining these tags pairwise in linear order, we arrive at `PPIS2.VVD`[4]., `VVD.II`, `II.MC` and `MC.UH` as the bigrams for this sentence.

(3) We$_{\texttt{PPIS2}}$ started$_{\texttt{VVD}}$ at$_{\texttt{II}}$ three$_{\texttt{MC}}$,, yes$_{\texttt{UH}}$ .. [DEV_005]

The total distribution of n-grams thus represents a model of syntagmatic relations between different kinds of word classes in a given text. Clearly, this is an incomplete model, as it can only capture adjacent dependencies. Nerbonne & Wiersma (2006) convincingly argue that this is not necessarily a severe problem, as simple measures often tend to correlate with more complex measures. They present a method for comparing two corpora based on the distribution of n-gram patterns. Their data source was a corpus of interviews with Finnish emigrants to Australia. Some of the informants emigrated as adults and some as children, and the goal of their analysis was to test for the influence of first-language interference in their spoken English. Sanders (2007; 2010) extends the general approach to syntactically parsed corpora. He surveys an extensive number of different methods for this, covering leaf-ancestor paths (Sampson 2000), i.e. the path from the root to each leaf node in a classic syntactic tree, as well as leaf-head and arc-head paths, a similar measure for dependency parses using either part-of-speech or dependency labels. Finally,

[3] Somewhat confusingly, both versions of the tagger and the corresponding tag set are named `claws` + number, although the form `CLAWS C7` is also used for the tag set `CLAWS7`. `CLAWS4 is the current version of the software.`

[4] I use a period as the character linking the individual POS tags in n-grams.

he includes two variants of counting the phrase structure rules observed in a particular tree. Comparing the results of these operationalizations, Sanders finds that "trigrams provide the most reliable results" (70), a fact that he explains by noting that the deeper syntactic measures require an additional parsing step, which increases the probability of wrong classifications[5].

The following is based on the exposition of this method in Nerbonne & Wiersma (2006)
.

Given a POS-tagged corpus consisting of two subcorpora, the analysis proceeds in the following way:

1. derive the n-grams from the subcorpora and count them
2. normalize the data using two normalization procedures
3. calculate the distance between the subcorpora, both per n-gram and aggregated over all n-grams
4. repeat the process using permuted versions of the original corpus to determine the reliability of both the per n-gram and total distances

Step 1 proceeds as described above, resulting in two per-subcorpus frequency vectors c^y and c^o, with the frequency-vector for the total corpus being their sum.

$$c^y =< c_1^y, c_2^y, ...c_n^y > N^y = \sum_{i=1}^{n} c_i^y$$
$$c^o =< c_1^o, c_2^o, ...c_n^o > N^o = \sum_{i=1}^{n} c_i^o$$
$$c =< c_1, c_2, ..., c_n > N = N^y + N^o$$

In step 2, the raw counts are transformed to correct for differing numbers of n-grams per subcorpus. This normalization procedure consists of two components. The first of these is the actual normalization. First, the raw frequencies are converted to relative

[5] The fact that part-of-speech tagging and syntactic parsing is usually done by automated, probability-based algorithms is somewhat troubling for further frequency-based analysis, due to a circularity: we want to determine which structures are frequent in a given text, and use frequency information from other texts to identify those structures. Automatic tagging does however lead to results that are reasonably close to those done by human annotators (cf. Sanders 2010: 71), and should thus be acceptable as a data source.

frequencies, i.e. each frequency is divided by the total number of n-grams per subcorpus.

$$f^y =< ..., f_i^y (= c_i^y / N^y), ... >$$
$$f^o =< ..., f_i^o (= c_i^o / N^o), ... >$$

Then, the relative frequencies are converted to per n-gram type proportions, by dividing the relative frequency of each type in each corpus by the sum of the individual per-type relative frequencies.

$$p^y =< ..., p_i^y (= \frac{f_i^y}{f_i^y + f_i^o}), ... >$$
$$p^o =< ..., p_i^o (= \frac{f_i^o}{f_i^y + f_i^o}), ... >$$

These can then be used to scale the original counts:

$$C^y =< ..., p_i^y * c_i, ... >$$
$$C^o =< ..., p_i^o * c_i, ... >$$

The end result are normalized frequency vectors that still contain the same amount of observations per n-gram type, but where the total amount of n-grams per subcorpus are more similar to each other than they are in the raw counts. This procedure should be applied several times, as n-grams that are more frequent in the smaller subcorpus do not have enough frequency mass. Iterating the process corrects for this. Nerbonne & Wiersma (2006) find that five iterations are enough to reduce the relative size difference to less than 0.1%, a result that is confirmed on the data discussed here.

The second normalization procedure is a simple scaling of the normalized frequencies by the average count of a given n-gram type. Let n be the number of n-gram types and N be the number of n-gram tokens. Then, the scaled, normalized frequency vectors can be calculated as:

$$s^y = C^y * 2n/N =< ...C_i^y * 2n/N >$$
$$s^o = C^o * 2n/N =< ...C_i^o * 2n/N >$$

The average of these vectors is 1. The reason for this step is to ease interpretation, and as it is a simple linear transformation its result can still be used to calculate linguistic distances in the next step.

Nerbonne & Wiersma (2006) provide several methods for step 3, the measuring of distances between the frequency vectors that result from the normalization procedure. Vector distance metrics such as the cosine distance frequently used in computational linguistics can be used, and two metrics based on the Recurrence metric by Kessler (2000) are proposed: *R*, the absolute difference of each n-gram to the average of both subcorpora, and *Rsq*, the same number squared. An aggregated distance over all n-grams can be calculated by summing the individual *R* and *Rsq* values, leading to the following formulas:

$$
\begin{aligned}
R &= \sum_i |c_i - \bar{c_i}| \\
Rsq &= \sum_i (c_i - \bar{c_i})^2 \\
\text{where } \bar{c_i} &= (c_i + c_i')/2
\end{aligned}
$$

The *R* metric is equivalent to the Manhattan distance (see Section 2.1.1) divided by two; in the interest of simplicity I employ the regular Manhattan distance here.

The final step is the evaluation of the per n-gram and total distances by means of permutation testing. The full original corpus is resampled without replacement into two new subcorpora, and steps 1 to 3 are applied to the new subcorpora. If the difference between subcorpora is meaningful with regard to a certain n-gram, we would expect a random subdivision to have a smaller distance than the original division. By repeating this process many times and counting the number of times where this assumption did not hold, we get a measure of the reliability both per n-gram and in total. If these counts are divided by the number of iterations, the results can be straightforwardly interpreted as significance values.

For dialectometric work, comparing only two variants is usually not enough. Thus, the method needs to be extended to work on more than two subcorpora. The simplest way of doing this is to simply apply the process to all pairwise combinations of subcorpora. As an addition, I propose a method that evaluates reliability by taking all of the corpus into account. To do this, the formulas given above need to be adapted. In most cases this is trivial, replacing the vectors for younger and older speakers by vectors for each subcorpus, but two steps operate on more than one vector and need to be replaced. The first is the conversion from relative frequencies to proportions, where the frequency needs to be divided by the sum of all frequencies of the n-gram type:

$$p^g =< ..., p_i^g (= \frac{f_i^g}{\sum_{e \in \text{subcorpora}} f_i^e}), ... > \text{ for each g } \in \text{ subcorpora}$$

Furthermore, in the second normalization, the number of subcorpora needs to replace the fixed number of groups:

$$s^g = C^g * |subcorpora| * n/N =< ..., c_i^g * \frac{|subcorpora| * n}{N}), ... >$$

Using this method results in a frequency matrix containing n-gram counts that are normalized and scaled to the total distribution of n-grams in the complete corpus. Permuting the whole corpus, I define the per n-gram reliability score q_i^g of each subcorpus as the sum of runs where the normalized count is higher than the original count, plus the number of runs where they are equal divided by two[6], divided by the number of runs. Calculating this for all subcorpora and all n-grams results in the reliability matrix Q. q_i^g is a directed measure: for a given subcorpus and n-gram, it will be close to 0 if that n-gram is reliably used more in that subcorpus compared to the whole corpus, close to 1 if it is underused, and close to 0.5 if it is used at a similar rate as in the total corpus. It can be translated into a unidirectional measure, indicating the extremeness of the distribution by means of the following formula: $p_i^g = 2 * \texttt{min}(q_i^g, 1 - q_i^g)$. Smaller values indicate more extreme distributions.

Finally, a method for identifying distinctive n-grams is required, both to reduce noise from n-grams that do not vary between dialects, and to aid interpretation and qualitative validation: a method that identifies known dialect features seems more trustworthy even on surprising results. I propose two distinctiveness metrics, one based on pairwise combination, one based on reliability scores. The first, labeled p-distinctiveness, simply counts the number of pairwise significant comparisons per n-gram; the higher the resulting number is, the more distinctive a given n-gram is. The second, labeled r-distinctiveness sums the reliability scores in their unidirectional formulation (i.e. p above); and a lower value here is interpreted as greater distinctiveness.

This leaves the question of how exactly to perform the permutation. In Nerbonne & Wiersma (2006), this was done on the basis of sentences: the corpus was divided into sentences, and those sentences were redistributed across subcorpora. Wiersma et al. (2011), however, recommend redistributing based on speakers, as sentences are not independent

[6] This is mainly relevant for n-grams that occur only rarely and in a small number of subcorpora, and where thus the count per subcorpus will be zero in most permutations

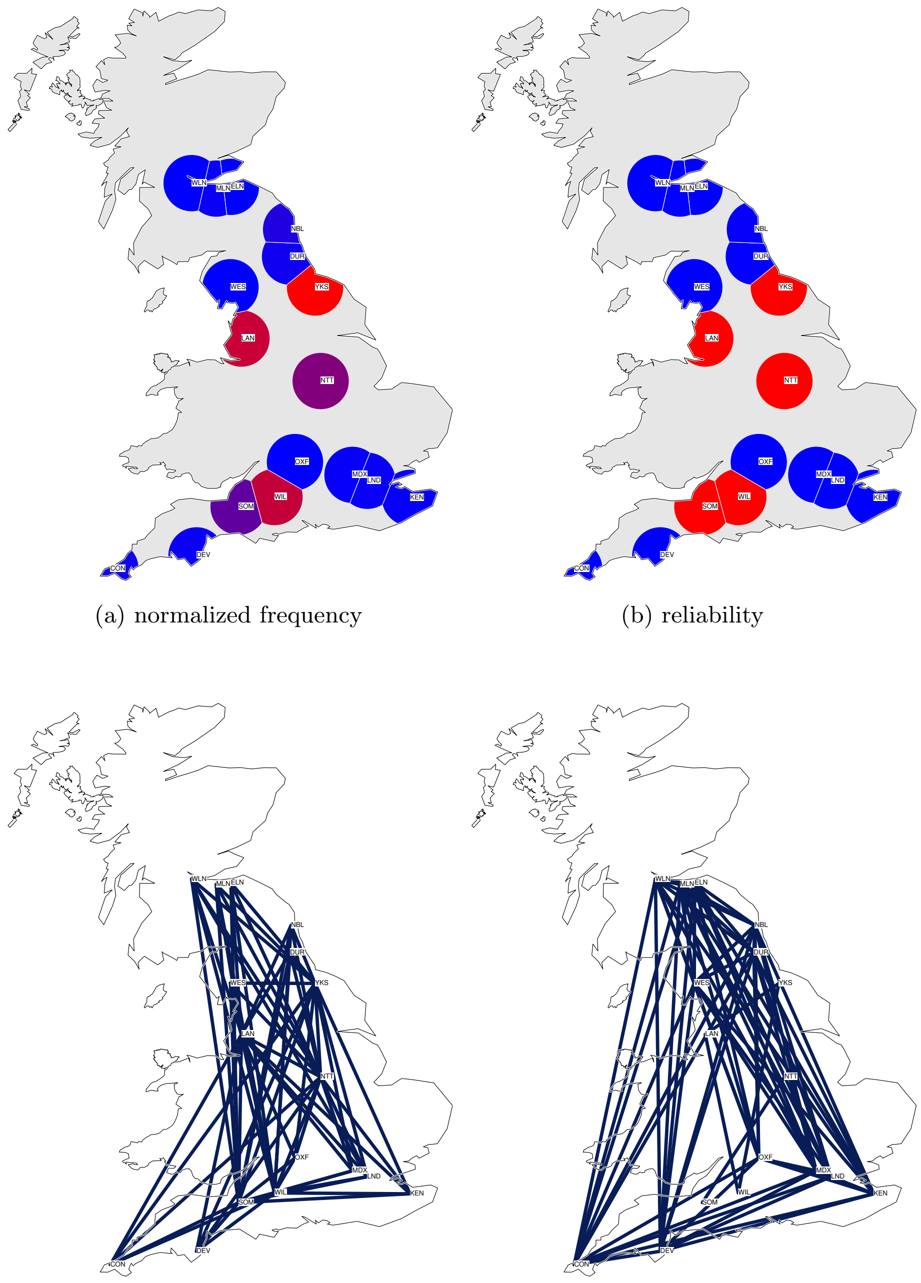
from one another: if an individual speaker has a preference for a certain pattern, it will appear in many sentences by that speaker, and thus bias the permutation toward finding a difference even if that speaker is not representative for the group. The most extreme difference of this is the case where in the comparison between two groups, a certain pattern appears only in material produced by one speaker, but that speaker uses it in many sentences. When resampling based on sentences, some sentences containing instances of this pattern will be likely to appear in both groups, and therefore the absolute difference with regard to that pattern will be less extreme. Resampling based on speakers, however, means that all instances of the pattern are moved as a whole and thus are always in exactly one group. The total difference is therefore always the same, and the permutation test would reject the significance of this difference. This, however, requires larger group sizes to find reliable differences. With low numbers of speakers in some groups, it is necessary to posit that the speakers are representative for the group; Sanders (2010: 31) therefore permutes his corpus based on sentences. For the geographic part of the analysis, I follow this approach, with the difference that, as sentence boundaries may be influenced by transcriber differences, whole conversational turns are resampled. For gender and age differences, the speakers per group are larger, and the more strict speaker-based permutation will be used.

Let me give an example for this process. Consider the case of the POS bigram `PPH1.VBDR`, *it were*. Table 3.5 displays the raw counts for this bigram in the column labeled *frequency*. The next column shows the normalized frequency values, i.e. the relative frequency of this bigram in the subcorpus compared to all other bigrams. Clearly, the frequency differences are very large overall. How do these results hold up to normalization across the whole corpus? The column *resampled norm* shows the result of an example run. The resulting numbers end up roughly normally distributed around the overall mean of 4.12. We now compare this value to that of the real data, and count it as 1 if the resampled run is larger and 0 if it is smaller. Through many repetitions, we end up with the values shown in the column *rel.* All values are very close to either one or zero, showing that the distributions throughout the whole corpus are rather extreme. Only two counties end up with a different result in any of the randomized runs. These are Northumberland and Somerset, whose frequencies are closest to the overall mean of 4.13. Using the formula above, we can determine the r-distinctiveness value of this bigram as 0.12; this is the most distinctive bigram in the corpus. Map 4a plots the resulting frequency distributions and Map 4b the inverted reliability score; in both maps, more reddish tones indicate that `PPH1.VBDR` is used more often.

To determine which individual distances are significant for this bigram, we permute

county	frequency	normalized	resampled norm.	rel. contr.	rel
CON	2	1.11	4.62	1	1.00
DEV	4	0.70	4.15	1	1.00
DUR	2	1.09	3.01	1	1.00
ELN	0	0.00	5.94	1	1.00
KEN	5	0.44	4.51	1	1.00
LAN	143	14.26	4.67	0	0.00
LND	3	0.57	3.81	1	1.00
MDX	1	0.47	4.19	1	1.00
MLN	1	0.69	3.96	1	1.00
NBL	4	2.15	4.77	1	0.96
NTT	93	9.29	4.09	0	0.00
OXF	0	0.00	4.38	1	1.00
SOM	32	6.91	3.27	0	0.02
WES	0	0.00	3.07	1	1.00
WIL	75	14.07	4.16	0	0.00
WLN	0	0.00	3.95	1	1.00
YKS	66	18.39	3.61	0	0.00

Table 3.5.: Bottom-up statistics for the bigram `PPH1.VBDR`, *it were*. Column *frequency* shows raw frequency per county, column *normalized* the results of the normalization process. Columns *resampled norm* and *rel. contr.* show the results of one permutation run: the normalized counts of a random corpus and the contribution of that run to the final score. The final column shows the reliability: values close to 1 show the original subcorpus is reliably smaller than expected based on random distribution, values close to 0 show the subcorpus is reliably larger.



(a) normalized frequency

(b) reliability

(c) significant differences

(d) non-significant differences

Map 4: Geographic distribution of the bigram `PPH1.VBDR`, *it were*, in FRED-S. In (a), blue represents low frequency and red high frequency. In (b), blue indicates a frequency reliably smaller than in a random corpus, red indicates one reliably larger. Bottom plots show lines between all counties that are pairwise significantly different (c) or not significantly different (d).

only between the two county subcorpora. For Devon and Somerset, as an example, the normalized values of the original corpus are 0.26 and 2.6[7]. Resampling this, we find, for example, normalized values of 1.2 and 1.6, a difference that is smaller than the original. As another example, consider Kent and London. We find normalized values of 0.19 and 0.25, while an exemplary permuted run leads to normalized values of 0.31 and 0.12. The difference in the permuted corpora is larger than the original one, and this run would therefore count against the significance of the difference between these two counties. Repeating this process for all corpus pairs a large number of times, we find that of the 136 pairwise combinations, 67 are significant at the $\alpha = 0.05$ level, or about half. Map 4c shows all combinations that lead to a significant result, while Map 4d shows those that are not significantly different. The pattern matches that in the frequency and reliability-based maps: pairs involving the high-frequency counties tend to be significant, while those between low-frequency counties usually are not.

In this work, I will restrict myself to uni- and bigrams. Trigrams are difficult to handle due to a large number of low-frequency patterns, and the results from an exploratory analysis showed that this measure is strongly affected by idiolectal differences. This is probably an effect of corpus size: Wiersma et al. (2011), who used trigrams, have a corpus about one third of the size of FRED-S and only compare two groups instead of seventeen. Increasing the amount of data, for example by supplementing FRED-S with additional texts from FRED, may make trigram analysis feasible in the future.

3.3. Chapter summary

This chapter introduced the data and methods used in the present work. The sources tapped here are the dialect corpus FRED and its part-of-speech tagged subset FRED-S. More specifically, the analysis based on FRED reuses the feature list of Szmrecsanyi (2013), but includes explicit model-building and represents those features that are intended to represent binary alternations as such. The analysis is also restricted to cases where the relevant sociolinguistic information about speakers is available. The analysis of FRED-S proceeds in a bottom-up fashion, determining dialectologically relevant features as part of the analysis process.

The discussion of methodology began with a summary of the CBDM methodology as it was used in Szmrecsanyi (2013). Two critical issues were identified: first, the method is likely to lead to less accurate results where little data is available, and therefore potentially

[7] The numeric values are different from those in Table 3.5, as there are fewer bigram types when only comparing two subcorpora. The ratio between them does not change.

overestimates the linguistic distances between groups in such cases. Furthermore, the method cannot take into account that sociolinguistic variation may matter. If some feature were to be used less often by younger speakers, for example, the linguistic distances between counties with differences in the average age would be inflated. As a solution to these problems, two ways of putting explicit probabilistic models were proposed: *mixed-effect models* using lmer and *generalized additive models*. Both alternatives proposed here allow for the inclusion of sociolinguistic predictors. It was shown that the normalization process is equivalent to a simple Poisson regression using *no pooling*, and that lmer modeling replaces this with *partial pooling*, where the strength of the evidence influences the values for individual counties. Where the evidence is weak, the per-county adjustments are pulled more toward the overall mean than where the evidence is strong. Using a simulation-based approach, normalization and mixed modeling were compared. It was shown that lmer leads, on average, to a better fit between model results and true values. This effect was most pronounced for rare features and low variability between groups. Generalized additive models, on the other hand, make a stronger assumption about geography by fitting a collection of functions to the geographic surface. Where the evidence is strong, abrupt transitions are possible, but where it is weak locations that are close together are more likely to end up similar to each other.

Finally, the methods of bottom-up analysis based on part-of-speech n-grams were introduced. They rely on syntagmatic relationships between word types in the dialect corpus, and estimate and quantify the difference between counties based on these. The methods proposed here rely on the permutation-based approaches used by Nerbonne & Wiersma (2006) and Sanders (2010), and are extended by introducing reliability and distinctiveness measures that compare individual counties to the overall distribution in the corpus.

The next chapter will discuss the results of this with regard to individual features or n-grams and their geographic and sociolinguistic patterns. The aggregational perspective can be found in Chapter 5.

4. Feature-based analyses

4.1. Model-based analyses

This chapter presents the results of the methods described in the previous chapter as they pertain to individual features and uni- or bigrams. First, the modeling results of both lmer models and GAMs will be presented in detail. Next, a case study will be used to investigate the effect of linguistically more sophisticated analysis. Synopsis sections will then survey the effects and distributions of the sociolinguistic factors, speaker age and gender, followed by a discussion of the most important geolinguistic patterns. I will then turn my attention to the bottom-up methods and describe a selection of n-grams that emerge as geolinguistically distinctive. This will be followed by a short investigation into the effect of gender and age on the bottom-up measures. A summary of the results of the bottom-up analysis will conclude this chapter.

4.1.1. Single feature models

This section proceeds as follows: I will present a brief description of each feature, and sketch the process by which the data was collected. This information is partially based on the feature descriptions in Szmrecsanyi (2013) and the extraction and coding protocols in Szmrecsanyi (2010a). Each feature will be illustrated with examples from the corpus. I will also present basic frequency information, including both the total number of observations and the number of speakers for whom the feature is attested. Here, it is important to note that this value can only give a lower limit: if 50 percent of all speakers use a certain feature, we know that it is available to them, but this does not mean that it is not available to the other half. Then, the results of the modeling process will be presented. For each feature, the reliable sociolinguistic predictors in both lmer and GAM models will be given, and the geographical results will be both described and projected onto a map. Finally, these results will be compared with the geographic distribution reported in the relevant literature. Where available, comparisons to the expert judgments in the recently compiled *World Atlas of Variation in English* (Kortmann & Lunkenheimer 2013, henceforth WAVE) will also be made. To reduce the potential for confusion between the numbering schemes, all WAVE feature numbers will be prefixed with ‘F’; for example, WAVE Feature 11 will be referred to as F11.

The individual maps should be read as follows: The background of the individual maps contains the topographic display of the GAM smoothers, as shown in Section 3.2.2.2; lighter and more red colors indicate higher frequencies while darker and more green colors indicate lower frequencies. Contour lines indicate the shape of the frequency distributions. In the interest of saving space, the county BLUPs will not be presented as individual tables, as they were in Section 3.2.2.1. Instead, they will be graphically projected onto the map, coloring each county indicator by their position in the range of attested BLUP values. The county with the lowest frequency adjustment will be colored in blue, and the one with highest in red. Intermediate points are in different shades of purple, with a redness that is proportional to the frequency adjustment.

4.1.1.1. Features 1–6: pronouns and determiners

4.1.1.1.1. Features 1 and 2: (non-)standard reflexives

These two features concern the number of either standard (Feature 1) or non-standard (Feature 2) forms of reflexives. Standard English reflexives follow an irregular paradigm, with *myself*, *yourself/yourselves* and *ourselves* using the possessive determiner + *-self/-selves* while others use the object form (*himself, themselves*) or an ambiguous form. Sentence (1a) shows a typical example for a standard reflexive. Some dialects regularize this pattern by allowing the other form (e.g. *hisself, theirselves*), as in (1b). Another possibility for forming a non-standard reflexive is a mismatch in number for plural reflexive pronouns, with speakers using singular *-self* with a plural pronoun or possessive determiner, as in (1c).

(1) a. Well I think it was because he perhaps went to school here himself, at Church School [...] [DEV_009]
b. He ’d forgotten you were coming or else he ’s have smartened hisself up. [WIL_022]
c. Put a banner across the road what they done theirself, Poor But Loyal. [LND_004]

According to WAVE (F11), some form of regularization in the reflexive paradigm is available in all dialect areas covered here. The feature is rated as rare in Welsh English and on the Isle of Man, and as pervasive in East Anglia and in the North. For the other regions, it is considered to be neither frequent nor rare.

The text frequencies for both features were determined automatically using a PERL script searching for the orthographic patterns above. Applying this on the texts in FRED with sufficient metadata yielded in total 1,099 standard reflexives and 146 non-standard reflexives; i.e. 11.7 percent of reflexive usages were non-standard. 73 speakers use a non-standard form at least once.

The two features are modeled in competition using logistic regression. The predicted odds are for non-standard realizations. In neither the lmer model nor the GAM do the sociolinguistic predictors or their interaction show a reliable effect.

In both models, the geographic predictors account for some of the variance: in the lmer model, the county random effect has a variance of 0.74; in the GAM, the geographic smoother is significant ($p < .001$). Map 5a displays the geographic results for both models, with the GAM smoothers in the background and the lmer intercept adjustments per county indicated via colored circles. In both models, the South of England and the Midlands show a higher probability of non-Standard reflexives, while Wales, the Scottish Highlands, the Isle of Man, Northumberland and especially the Hebrides have a lower probability. This matches the classifications in FRED, especially as far as East Anglia, the Isle of Man, and Wales are concerned. Only the pervasiveness of non-standard reflexives in the North of England is not apparent from the plot. The GAM explains 34.4 percent of the deviance.

4.1.1.1.2. Features 3 and 4: archaic pronouns *thee, thou, thy* and *ye*

Another source of variation concerns the archaic forms of the second person pronouns. Feature 3 concerns itself with forms of the archaic second person singular pronoun, *thee, thou*, and *thy*, which originally covered all singular uses and later became restricted to informal address. These forms are no longer in general use in Standard English, where they are restricted to specific contexts such as religious language. Nevertheless, they are still available in some dialects, especially in the North of England (Evans 1969). Example (2a) shows a typical case. Many instances, however, are direct quotations of past utterances (2b) or of religious material (2c). Feature 4 covers the originally plural nominative form *ye*, which came to be used as the singular form in formal contexts before being replaced by the object form *you* during the sixteenth century (Raumolin-Brunberg 2005). Like *thou*, it is still available in some dialects (2d). Wales (2006: 181ff and references therein) notes that *thou* is still in use in the North of England, and *ye* in Northumberland.

(2) a. Ah thou 'll know there 's never been any trouble [ANS_004]
b. [. . .] and they used to say, Is thee for hiring lass? [LAN_002]
c. And a Christian is taught, thou shalt not kill, irrespective. [LAN_012]

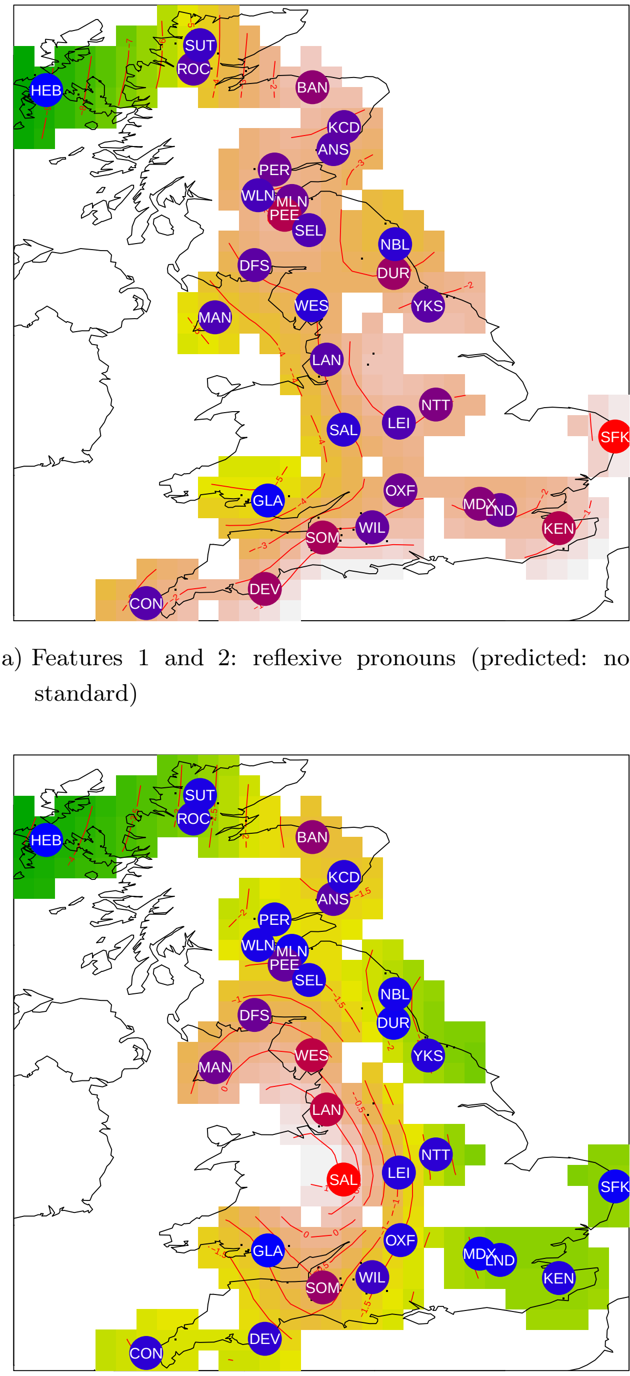
d. [...] did ye see her face, she was going to hit us all. [ELN_011]

WAVE covers this variation as part of a broader feature (F35) concerning second person singular pronouns other than *you.* This is rated as pervasive in Scottish English, neither frequent nor rare in the North and the Southwest of England, and as rare in Wales.

The text frequencies for these features were determined automatically using a PERL script searching for the respective words as lexical strings. Applying this on the texts in FRED with sufficient metadata yielded in total 172 instances of *thee, thou, thy*, used at least once by 42 speakers, and 234 instances of *ye*, used by 51 speakers.

The archaic pronouns *thee, thou, thy* (Feature 3) and *ye* (Feature 4) are analyzed using a count-based model. The lmer models detect sociolinguistic effects for both features: for Feature 3, gender, age and their interaction influence the frequency of usage. Women and, surprisingly, older speakers use the archaic form less often. The interaction of both indicates that the gender difference is less pronounced for older speakers. The effect of gender is not unexpected: as Wales (2006: 185) reports, in several communities it was found that male speakers use *thou* in more contexts, “as a sign of kinship and paternity, or of camaraderie in the local pub or club”. She also notes an article from the *Guardian* reporting a spreading use of *thou* among children in 1983; this should, however, not affect the data in FRED as the included speakers are much older. For Feature 4, only age has a reliable effect, such that older speakers use the archaic form more often. The GAMs are more conservative and do not detect an effect of any sociolinguistic predictor for Feature 3, although there is a non-significant trend for gender ($p < .12$). With the exception of age, the numeric values have the same effect direction as in the lmer model. For Feature 4, there are again no significant effects, but numeric values have the same effect direction as in the lmer model. Taken together, there is weak evidence for a gender difference in the use of *thee, thou, thy* and for an effect of speaker age in the use of *ye.*

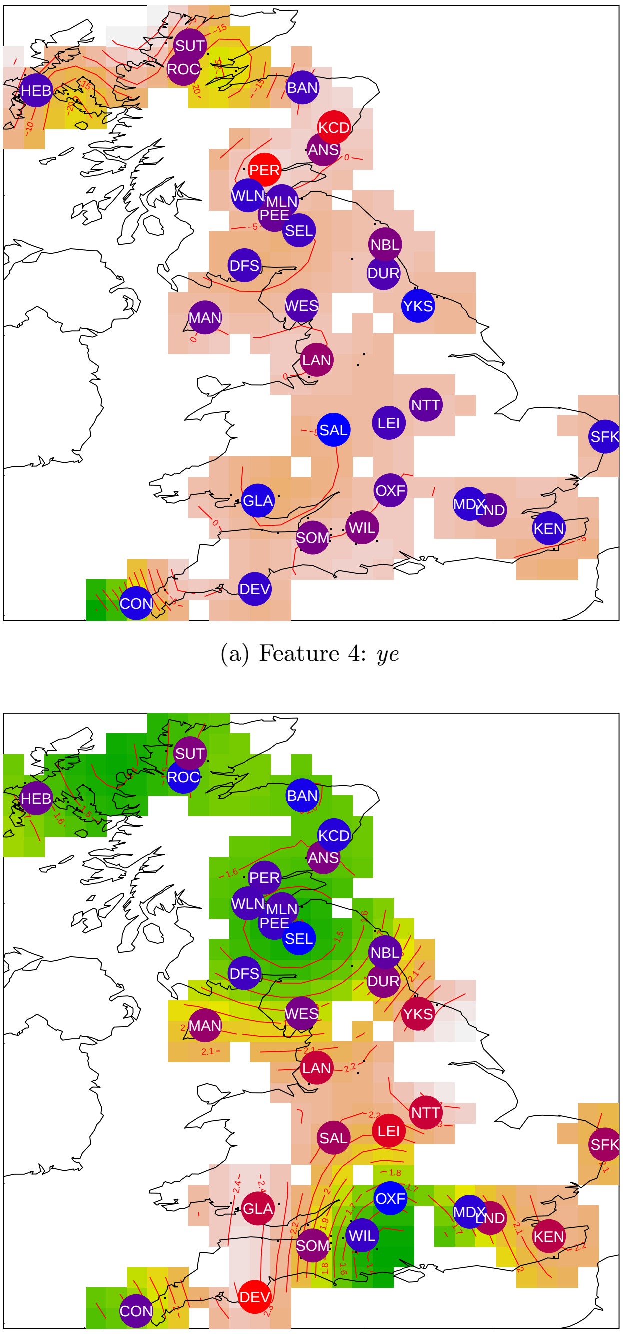
Map 5b displays the geographic distribution of Feature 3 according to the lmer model and the GAM. In both, geography accounts for part of the variance, with the lmer county random effect having a variance of 2 and the geographic smoother in the GAM being highly reliable ($p < .001$). *Thee, thou*, And *thy* are particularly frequent in the Western Midlands and North, and particularly infrequent in the Southeast of England, the Scottish Highlands, and the Hebrides. Map 6a shows the geographic distribution of Feature 4. Again, geography has a marked effect, with the lmer county random effect having a variance of 4.03 and the GAM smoother being highly significant ($p < .001$). *ye* is especially frequent in the Scottish Lowlands, and infrequent in the very Southeast and rare in Cornwall.



(a) Features 1 and 2: reflexive pronouns (predicted: non-standard)

(b) Feature 3: *thee, thou, thy*

Map 5: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).



(a) Feature 4: *ye*

(b) Feature 5: *us*

Map 6: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring).More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

The GAM for Feature 3 explains 40.3 percent of the deviance, and the GAM for Feature 4 is even better at 67.6 percent.

4.1.1.1.3. Feature 5: *us*

The first person plural object pronoun form *us* can be used in several non-standard ways: first, some dialects allow *us* as the subject form, as in (3a). In WAVE, this is part of a general feature concerning object pronouns in subject function (F31), which is neither frequent nor rare in the Southwest, and rare in Wales and the North of England. Far more dialects allow the use of *us* + noun phrase as a subject, as in (3b), which is rated as neither frequent nor rare in Scotland, the North, and the Southwest of England, and frequent everywhere else. *Us* can also be used as a possessive determiner, as in (3c). In WAVE, this (F27) is considered neither frequent nor rare in the North of England, rare in Wales, and absent in all other regions.

(3) a. [...] us had to walk there and walk back, winter and summer. [DEV_010]
b. And us boys used to have to ride 'em, [...] [KEN_002]
c. [...] when we got in us teens we used to have to help us mother with her cleaning [...] [NTT_006]

The text frequencies for this feature were determined automatically using a PERL script searching for tokens of *us*. Note that this feature counts all instances of *us*, including standard usages. Applying this on the texts in FRED with sufficient metadata yielded in total 2,907 instances of *us*, and 249 speakers used it at least once.

Feature 5 was analyzed using a count-based model. The lmer model detects a sociolinguistic effect for gender, such that female speakers use *us* more often than male speakers do. There is also an interaction of gender and age, such that the difference between women and men is even more pronounced for older speakers. The GAM confirms only the gender difference; while all coefficients point into the same direction, neither age nor the interaction of gender and age emerge as significant in the GAM.

Us again shows geographic differences in its frequency distribution, although it is less pronounced than for the previous features. The lmer county random effect has a variance of 0.21, and the GAM geographic smoother is significant ($p < .01$). This feature is most frequent in the Midlands and the lower part of the North of England, and particularly rare in Scotland and in the central regions of the English South. Again, the results mostly fit the WAVE classifications, complicated by the fact that Feature 5 indirectly maps to several WAVE features: Scotland, which has none of these features, shows very low usage of *us*, while intermediate regions such as the North and the Southwest of England are rated lower on some WAVE features and higher on others. The result is also consistent with other dialectological work, as was discussed in Section 3.2.2.2. The GAM explains 23.4 percent of the Deviance.

4.1.1.1.4. Feature 6: *them*

One of the most pervasive dialect features involves usage of the third person plural object pronoun *them* as a demonstrative with plural nouns, as in (4a). In WAVE, where this is F68, it is classified as pervasive in all relevant regions except for Scottish English and the Southwest, where it is considered neither frequent nor rare.

(4) a. [...] they didn't like you, them blokes what worked there.
    b. In St. Ives they called them troys. [CON_002]

The text frequencies for this feature were determined automatically using a PERL script searching for instances of *them* followed by a word ending in -*s*[1]. The count thus presents an upper boundary for the pervasiveness of this feature, as standard uses of them as in (4b) are also included. Applying this on the texts in FRED with sufficient metadata yielded in total 687 instances of *them*, and 160 speakers used it at least once.

Feature 6 was analyzed using a count-based model. The lmer model detect a significant effects for gender, with women using the potentially non-standard form significantly less often. The GAM is again more conservative, and only detects a marginally significant effect for gender ($p < .06$) in same direction as the lmer model.

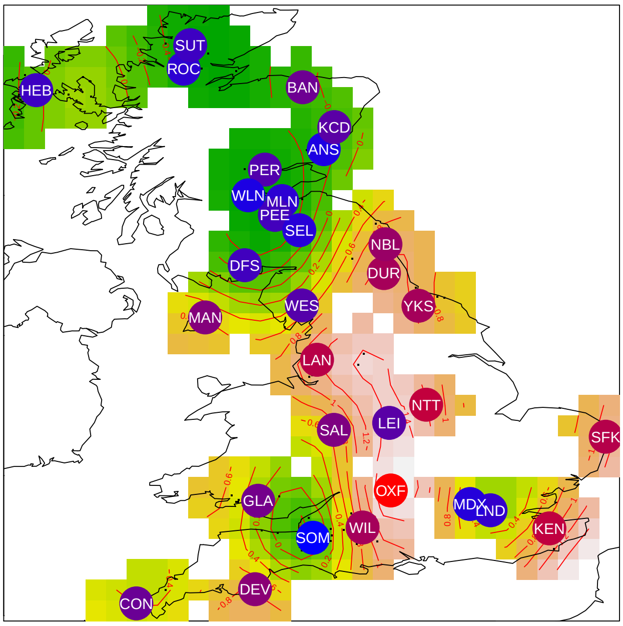
The lmer county random effect has a variance of 0.43, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 7 illustrates these distributions: *them* is frequent in the Midlands, the North of England, the very Southeast, and parts of the Southwest; it is particularly infrequent in Somerset and Scotland. Here, the match between WAVE classifications and modeled results is particularly nice: the two areas with the lowest frequency are also ranked lower in WAVE. The GAM explains 23 percent of the deviance.

4.1.1.2. Features 7–12: the noun phrase

4.1.1.2.1. Feature 7: synthetic adjective comparison

For some adjectives, English allows two variants of the comparative: a synthetic version consisting of the adjective suffixed by -*er*, as in (5a), and an analytic version consisting

[1] Some high-frequency words ending in -*s* that are not plural nouns, such as *as*, were also excluded. See Szmrecsanyi (2010a: 10) for the complete list



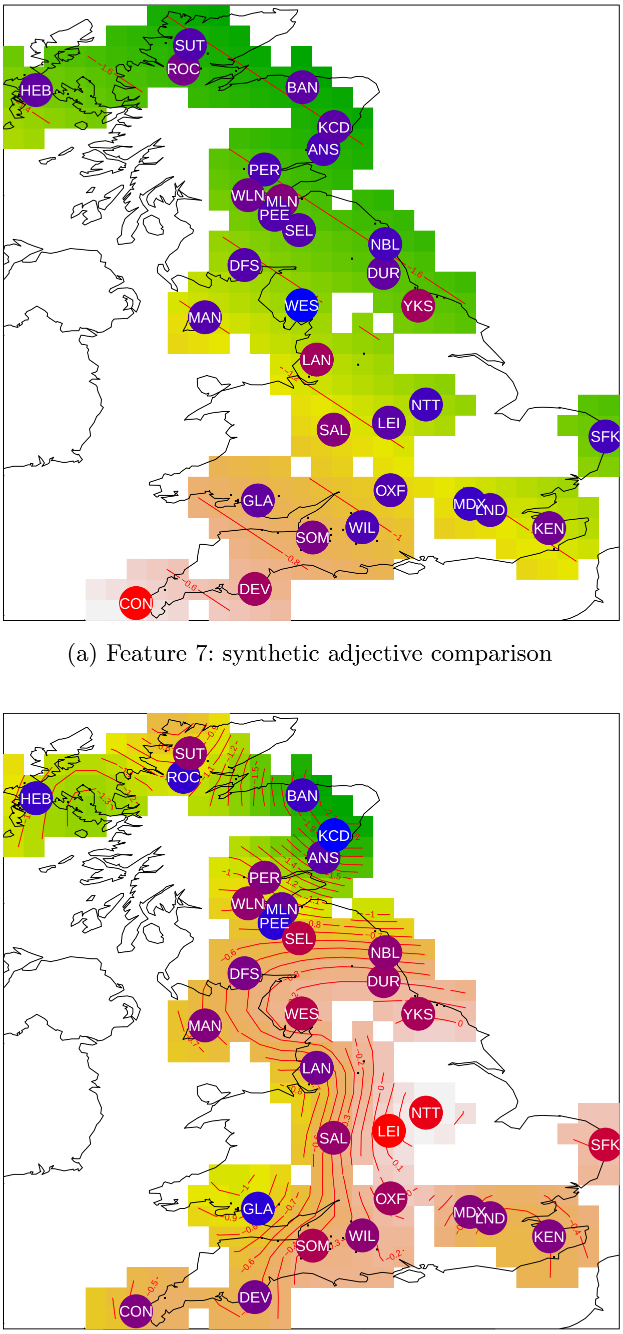
Map 7: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring) for *them*. More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

of the adjective preceded by *more*, as in (5b).

(5) a. That was a cheaper way of buying it for them. [WES_001]
b. [. . .] and she was more free than us, you see [. . .] [LAN_001]

The text frequencies for this feature were determined automatically using a PERL script searching for a list of adjectives that can appear with either synthetic or analytic comparison and were suffixed by `-er`. the texts in FRED with sufficient metadata yielded in total 96 instances of synthetic adjective comparison, with 57 speakers using it at least once.

Feature 7 shows no sociolinguistic effects in either lmer model or GAM. The geographic pattern is very weak, with an lmer county random effect variance of 0.12 and only a marginally significant geographic smoother in the GAM ($p < 0.06$). Map 8a displays this pattern, with a high frequency in Cornwall and lower values in the other counties, particularly in the North. The model quality here is bad as well, explaining only 4.5 percent of the deviance.



(a) Feature 7: synthetic adjective comparison

(b) Features 8 & 9: genitive alternation (predicted: *s*-genitive)

Map 8: Geographic effects in the lmer model (dot coloring) and GAM (area coloring) More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).

4.1.1.2.2. Features 8 and 9: the genitive alternation

English has two ways of realizing the genitive: the *of*-genitive, in which the possessor is realized as a prepositional phrase following the possessum, and the *s*-genitive, in which the possessor precedes the possessum and is marked using the clitic *'s*. Both forms have existed since the later stages of Old English, and their prevalence has shifted quite dramatically over time (see Wolk et al. 2013: Section 2 for a review). While this variable has not received a lot of attention from the dialectological perspective, differences between British and American English concerning the determinants of this choice are well-studied (e.g. Szmrecsanyi et al. 2014 and references therein).

Candidates for this alternation were identified automatically using a PERL script searching for *of* and *'s*, then manually screened to remove instances that were not interchangeable genitives. Applying this process on the texts in FRED with sufficient metadata yielded in total 1,255 instances of the *of*-genitive, and 226 speakers used it at least once. For the *s*-genitive, there are 971 tokens produced by 217 speakers. Overall, 43.6 percent of tokens are *s*-genitives.

Features 8 and 9 are modeled in competition using logistic regression. The predicted odds are for the *s*-genitive. Both the lmer model and the GAM show similar, very reliable effects of gender: female speakers are more likely to use the *s*-genitive than male speakers ($p < .001$). Neither age nor its interaction with gender are significant in either model. Concerning geographic effects, the lmer county random effect has a relatively small variance of 0.19, but the GAM smoother is very reliable at $p < .001$. Map 8b shows the pattern: the *s*-genitive is frequent throughout England and rarer in Wales and in Scotland, particularly in the northeast of the Scottish Lowlands. The GAM explains 16.6 percent of the deviance.

4.1.1.2.3. Feature 10: preposition stranding

English has two options when *wh*-moving the complement of a preposition: the preposition can be moved to the front of the *wh*-marker as in (6a), a process called pied piping, or it can be left at its original spot, called preposition stranding, as in (6b). Herrmann (2003: 124) finds that dialect speakers in her corpus data clearly prefer preposition stranding over pied piping for relative clauses, both in total by over 90 percent and in those cases where pied piping could have easily occurred (i.e. without changing the relative marker) by over 60 percent. Even in sentence (6b) where the preposition is moved, it is reinserted at it original place and combined with a resumptive pronoun.

The text frequencies for this feature were determined by manually screening a list of instances that fit the general pattern of preposition stranding, which in turn was extracted from the corpus automatically using a PERL script. Applying this process on the texts in FRED with sufficient metadata yielded in total 747 instances of preposition stranding, with 223 speakers using it at least once.

(6) a. You could get over whatever you were preparing, for whichever crop you were preparing, you could get over it with the agricultural implements. [CON_010]
b. The rate of output is about two thousand two hundred pounds of steam per hour, of which we use about three quarters of it to generate our jam pans. [LAN_019]

Feature 10 shows neither effects of sociolinguistic predictors nor of geography (lmer county random effect variance = 0, GAM geographic smoother significance > 0.34).

4.1.1.2.4. Features 11 and 12: cardinal number + *year(s)*

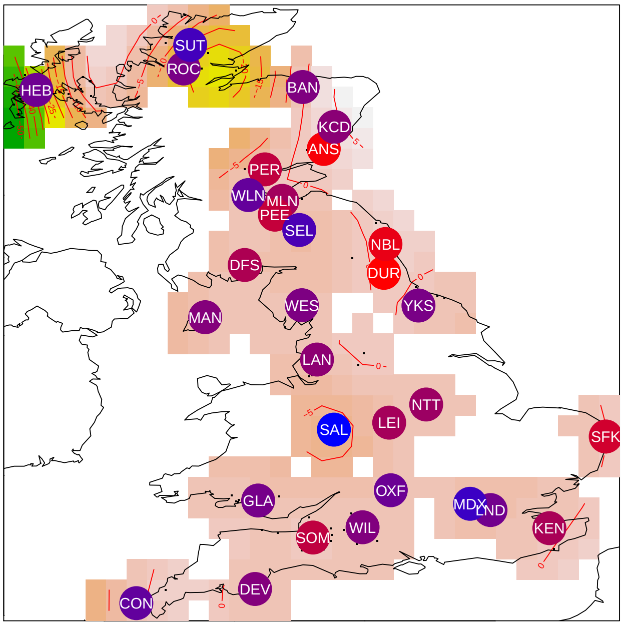
In some dialects of English, the plural marking on *years* in contexts like (7a) and (7b) is optional. WAVE subsumes this under a more general feature concerning absence of plural marking after quantifiers (F56), which is judged as pervasive in East Anglia, absent in Scottish English and the Isle of Man, and neither frequent nor rare in all other relevant regions.

(7) a. Yes, six year, seven year. [DEV_002]
b. And uh, I took to that, farming, and I stuck it for two years. [DEV_002]

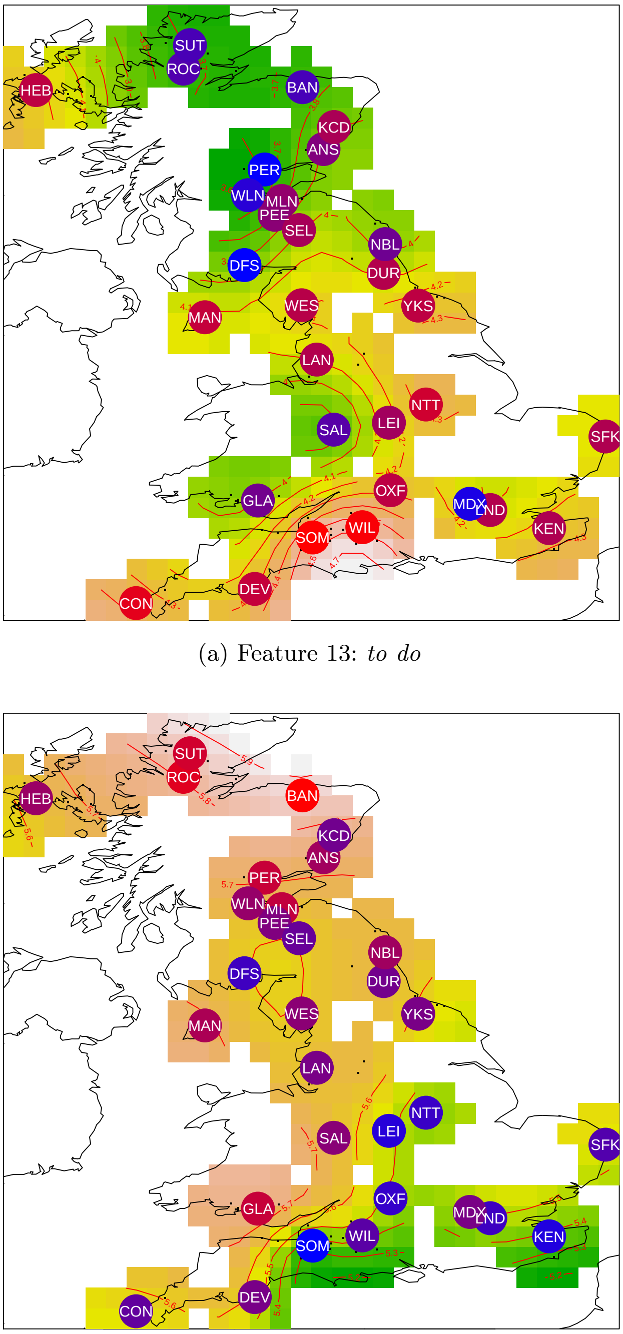
The text frequencies for both features were determined automatically using a PERL script searching for the orthographic patterns `year` and `years` preceded by a number word. Applying this on the texts in FRED with sufficient metadata yielded in total 1,018 instances of *years* and 351 of *year*; i.e. 25.6 percent of the total instances use *year*. 108 speakers use *year* at least once, compared with 219 for *years*.

Features 11 and 12 are modeled in competition using logistic regression. The predicted odds are for the version without *-s*. In the lmer model, the sociolinguistic predictors have no effects. Here, the GAM is less conservative, and finds a gender difference, such that female speakers use the non-standard variant *year* less often ($p < .05$).

In both models geography has an effect, with the lmer county random effect having a variance of 1.39, and the GAM geographic smoother being highly significant ($p < .001$). The geographic distribution is quite extreme, as Map 9 shows: *year* is particularly probable in the east of the North of England and part of the Scottish Lowlands, and very unlikely



Map 9: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring) for cardinal number + *year(s)* (predicted: *year*). More red dots and areas indicate higher odds for the predicted realization, more blue dots and green areas indicate lower odds.



(a) Feature 13: *to do*

(b) Feature 14: *to be*

Map 10: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

in Shropshire and the western part of the Hebrides. The other counties show intermediate probabilities. The GAM explains 26.2 percent of the deviance.

4.1.1.3. Features 13–16: primary verbs

These three features concern themselves with the primary verbs of English, i.e. the verbs that can function both as auxiliaries and as main verbs. All of them are involved in several dialectal features, and extended usage contexts for these verbs should be reflected in increased frequencies.

The text frequencies for these features were determined automatically using a PERL script by searching for the orthographic verb forms of the three verbs, including contractions and non-standard forms. Applying this on the texts in FRED with sufficient metadata yielded in total 21,899 instances of *to do*, 80,701 instances of *to be*, and 28,519 instances of *to have*. Unsurprisingly, almost all speakers use all three primary verbs: 271 speakers use *to do*, 273 *to be*, and 270 *to have.*

4.1.1.3.1. Feature 13: *to do*

Feature 13, the primary verb *to do*, is modeled using counts. The lmer model results in significant effects for gender and the interaction of gender and age (both $p > .001$), such that female speakers use *to do* more often, with this difference decreasing with age. The GAM confirms the effect for gender ($p > .001$) but not the interaction, although the numeric coefficient points into the same direction.

The lmer county random effect only shows a slight geographic distribution, with the variance being 0.09. The geographic smoother term in the GAM, however, is highly significant ($p < .001$). Map 10a illustrates this distribution. *To do* is particularly frequent in the Southwest of England, and more rare in Scotland, Wales, and Shropshire. Middlesex and its neighboring county London exhibit very different behavior in lmer and GAM models; the lmer model finds a large difference between the two, while the GAM only finds support for a rather small difference. The GAM explains 42.5 percent of the deviance.

4.1.1.3.2. Feature 14: *to be*

Feature 14, the primary verb *to be*, was analyzed using a count-based model. Neither the lmer model nor the GAM find evidence for an influence of sociolinguistic predictors. Geography, however, does seem to have an effect: while the lmer county random effect variance is very small at 0.02, the geographic smoother term is highly significant ($p < .001$) Map 10b illustrates this, with the South of England having lower frequencies of *to be* while the North of England as well as Scotland have higher frequencies. The GAM explains 32.3 percent of the deviance.

4.1.1.3.3. Feature 15: *to have*

Feature 15, the primary verb *to have*, was analyzed using a count-based model. The lmer model finds significant effects for both sociolinguistic predictors and their interaction, such that female speakers use *to have* more often while older men have lower frequencies, and for older women the frequency difference is even larger. The GAM only supports the gender difference ($p < .001$) and not the age effect or the interaction, although the effect directions are the same.

As with *to be*, *to have* shows little variance (0.03) as an lmer county random effect, but is highly significant as a geographic smoother in the GAM ($p < .001$). Map 11a shows the distribution: high frequencies can be found in England, with the exception of Middlesex and London, while Scotland has predominantly low frequencies. The GAM explains 23.4 percent of the deviance.

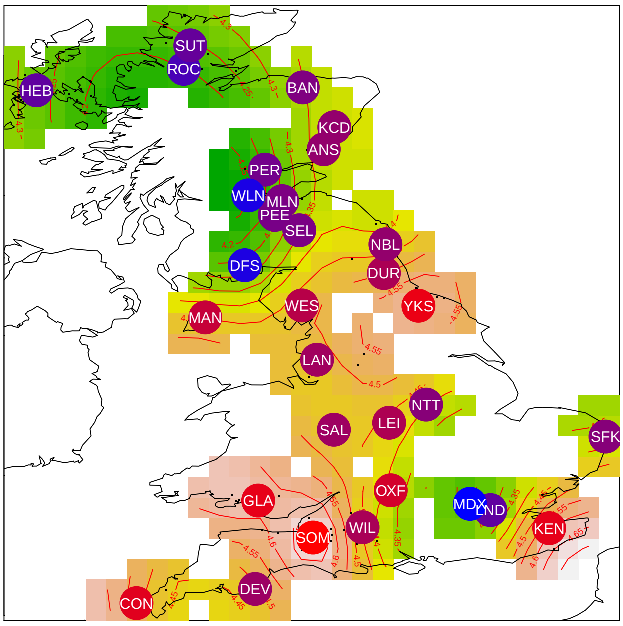
4.1.1.3.4. Feature 16: marking of possession: *have got*

To indicate possession in British English, both the primary verb *have* (8a) and forms of *have got* (8b) can be used. The latter is a relatively recent form: Denison (1993: 341) places its development in the Modern English period, and Schulz (2012: 120), in her detailed discussion of the phenomenon and its history, notes a lot of linguistic interest in the phenomenon since its first appearance in a dictionary in 1773.

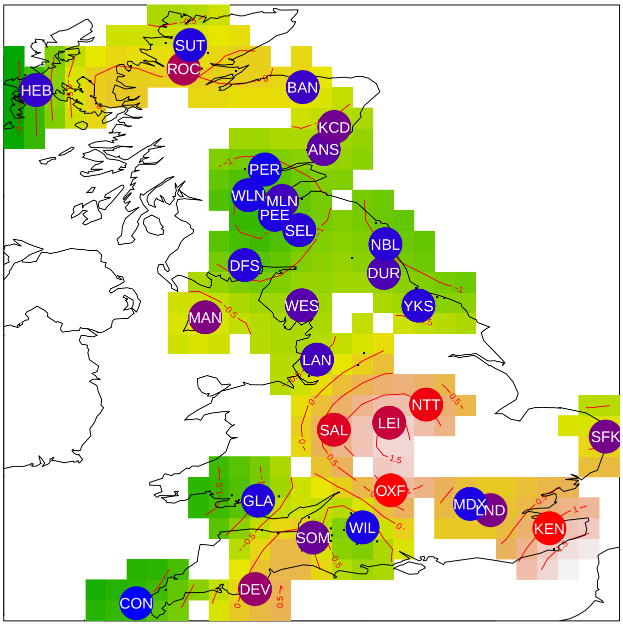
(8) a. Oh aye, well she has a good job. [WES_014]
b. Louise has got a car. [ELN_008]

The text frequencies for this feature were determined automatically using a PERL script searching for forms of *have* followed by *got*. Instances followed by *to*, indicating obligation instead of possession, as well as clear cases of particle verbs involving *get* were excluded. Applying this on the texts in FRED with sufficient metadata yielded in total 366 instances of *have got*, and 128 speakers used it at least once.

Feature 16 was analyzed using a count-based model. No sociolinguistic predictors have an effect in any of the models. Geographic variability is somewhat high, with an lmer county random effect variance of 0.83 and a GAM geographic smoother significance at the level of .001. Map 11b illustrates this distribution, with the Midlands and Kent having high frequencies and Cornwall, the North of England, the southern Scottish Lowlands,

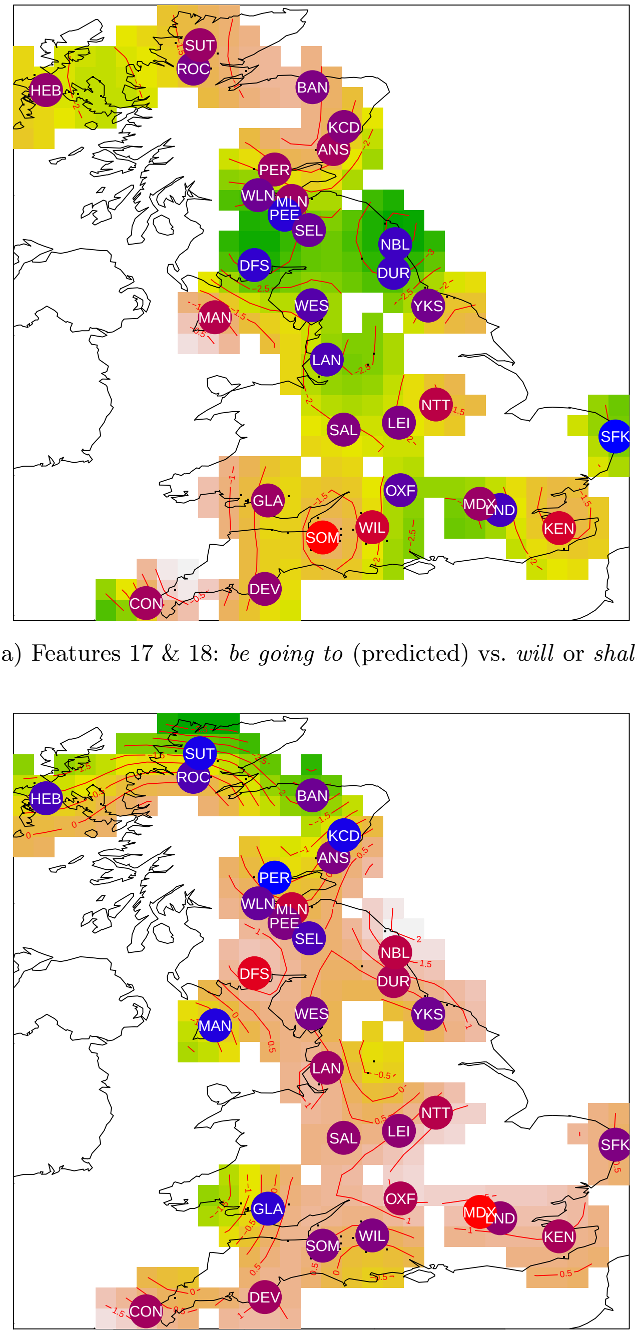


(a) Feature 15: *to have*



(b) Feature 16: *have got*

Map 11: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.



(a) Features 17 & 18: *be going to* (predicted) vs. *will* or *shall*.

(b) Features 19 & 20: *would* vs. *used to* (predicted)

Map 12: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher odds for the predicted realization, more blue dots and green areas indicate lower odds.

Wales and the Hebrides using *have got* rarely. The GAM explains 43.9 percent of the deviance.

4.1.1.4. Features 17–23: tense and aspect

4.1.1.4.1. Features 17 and 18: future markers *be going to* and *will* or *shall*

The development of *be going to* as a future marker is one of the textbook examples of grammaticalization (cf. Hopper & Traugott 2003: 88f.). The older future forms *will*, as in (9b), and *shall*, as in (9c), emerged during late Old English and became fully productive and frequent during the Middle English period (Poplack & Tagliamonte 1999: 318). They are still in productive use today. *Be going to* typically carries intentional meaning, as in (9a). Mair (2004: 128) reports a first example of *be going to* as a future marker in the *Oxford English Dictionary* dating back to 1482 and that the grammaticalization process "was complete by the end of the 17th century", but also that usage frequencies remained relatively low until a major surge at the beginning of the twentieth century.

(9) a. I 'm not going to get rid of any of it. [WIL_024]
b. [...] a furniture sale will be held in the Assembly Rooms tomorrow starting at eleven a-m. [WES_013]
c. I shall be eighty-four in February that is 1876 isn't it? [WES_001]

Instances of both constructions were selected automatically by searching for orthographic strings matching the forms of *be going to*, *will*, and *shall*, and in cases where non-future usages were probable those were manually screened. Applying this process on the texts in FRED with sufficient metadata yielded in total 515 instances of *be going to* and 3627 of *will* or *shall*, i.e. the overall percentage of *be going to* is 12.4. 158 speakers use the first at least once, and 242 speakers the second.

Features 17 and 18 are modeled in competition by means of logistic regression. The predicted odds are for *be going to*. Both the lmer model and the GAM show the same effect: older speakers use *be going to* less often to mark the future. For the lmer model, this effect is marginally significant, while the effect in the GAM is reliable ($p < .01$). This replicates Tagliamonte et al. (2014), a recent study of the ongoing grammaticalization of *going to* in British dialects.

Geographical variation in the choice of future marker is present in both models. The lmer county random effect has a variance of 0.13, and the GAM smoother has a highly significant effect ($p < .001$). As can be seen in Map 12a, *be going to* is used particularly often in the Southwest, particularly in Somerset, while the North of England and the southern Scottish Lowlands prefer *will* or *shall.* The GAM explains 22.7 percent of the deviance.

4.1.1.4.2. Features 19 and 20: habitual past: *would* or *used to*

There are two ways in English to express past habituality: one consisting of *used to* + VERB as in (10a), the other of *would* + VERB. Both are in principle roughly interchangeable, and researchers often analyze them in competition (Schulz 2012, Tagliamonte & Lawrence 2000)

(10) a. Now we used to do everything for the customer. [WES_023]
b. <IntER> Did you do the shopping?
<u Lang1p> Oh, I would go on errands, yes, and you go to, you go on errands for other people. [LAN_003]

Instances matching either marker were extracted automatically, and instances of *would* were then manually screened to ensure interchangeability with *used to.* Applying this on the texts in FRED with sufficient metadata yielded in total 1,845 instances of habitual *would* and 3,420 of *used to*, or 35 percent *would.* 236 speakers use *would* at least once compared with 252 for *used to.*

Features 19 and 20 are modeled in competition by means of logistic regression. The lmer model finds significant effects of both gender and age, such that women and older speakers are more likely to use *used to.* The GAM confirms this, as both of these effects are significant and have the same direction.

There is considerable geographic variation: the lmer county random effect has a variance of 1.11, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 12b illustrates these distributions. *Used to* is most frequent in Middlesex and generally in England and the south of Scotland. Wales, the Isle of Man, and the north of Scotland use *would* more often. The GAM explains 26.2 percent of the deviance.

4.1.1.4.3. Feature 21: progressive verb forms

Forms similar to the modern progressive (as in (11)) have arguably existed in English since the Old English period, although their functions were not quite the same (cf. Denison 1993: 371, 381). As Hundt (2004: 47) notes, “the rules for the use of this aspectual form as we know them today only emerge in the course of the seventeenth century”, and frequency of the progressive begins to increase dramatically during the nineteenth century.

(11) [...] and poor Milly was sitting there and she stuck her fork in the jam tart [...] [YKS_008]

WAVE contains two features relevant to progressives, both concerning extensions of their use: the first, F88 concerns itself with progressives for stative verbs such as *like* or *want*, which is rated as frequent on the Isle of Man and in Scotland, and as neither frequent nor rare in the North of England and in Wales. F89 covers the extension to habitual contexts, which is frequent only on the Isle of Man and neither frequent nor rare in Wales. Such extended uses should, everything else being equal, lead to increased frequencies.

The text frequencies for this feature were determined automatically using a PERL script. Applying this on the texts in FRED with sufficient metadata yielded in total 1,280 instances of progressive verb forms, and 244 speakers used it at least once.

Feature 21, progressive verb forms, was analyzed using a count-based model. The lmer model finds a significant effect for gender, age, and their interaction: women use progressive verb forms less often while older speakers use them more frequently, and the gender difference decreases for older women. The GAM does not confirm this, although the coefficients keep their numeric direction.

There is some geographic variability, with the lmer county random effect variance being 0.3, and a highly significant ($p < .001$) geographic smoother in the GAM. Map 13a visualizes this. Progressive verb forms are particularly frequent in the Hebrides, and show a combination of east/west and north/south distributions, such that more eastern and more southern dialects use fewer progressive verb forms, with the very Southeast of England having the lowest frequency. East Anglia is an outlier, exhibiting intermediate frequencies despite its very eastern location. These results are somewhat consistent with WAVE, especially where the Isle of Man and Wales are concerned, with the picture being less clear for the North and Scotland. The GAM explains 34.5 percent of the deviance.

4.1.1.4.4. Features 22 and 23: present perfect: auxiliaries *be* and *have*

Be is the original present perfect auxiliary in the Germanic languages, although the perfect with *have* was also already available in Old English (Denison 1993: 344, 346). Denison considers the effective complete replacement of *be* to have happened during the nineteenth century (1993: 395), but in some dialects the older form is still available today. While the majority of cases in FRED use the standard auxiliary as in (12b), many instances of the present perfect with *be* as in (12a) can be found. WAVE covers this feature as F102, "*be* as a perfect auxiliary". It is rated as neither frequent nor rare on the Isle of Man and in Scotland, and as rare in the North and the Southwest of England.

(12) a. As you can see, we use it also, we 're re-roofed it and we use it for package and storage [...] [LAN_019]
b. No I haven't sold the calves yet, I 'll sell them next month. [HEB_019]

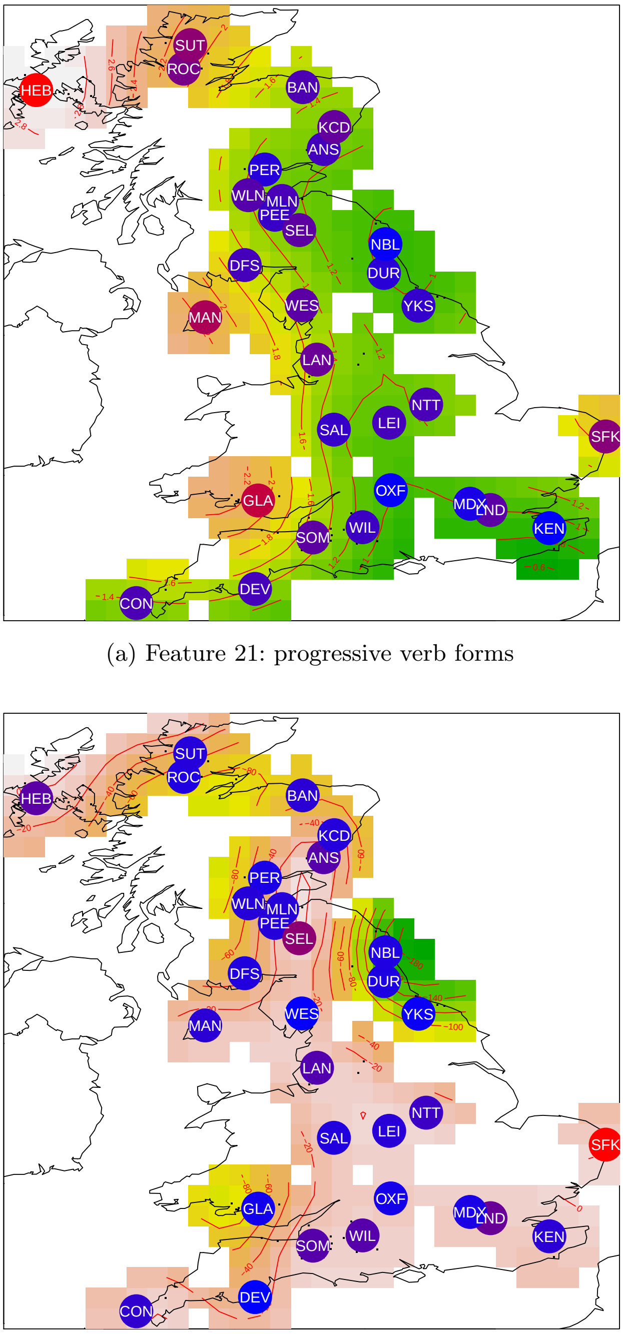
The text frequencies for these features were determined using a two-step process: first, all instances of forms of *be* and *have* were extracted, ignoring those contexts where a present perfect can be safely ruled out. The remaining cases were then manually screened. Due to the large number of tokens, for *have* only the first 1500 words of each text were included. Applying this on the texts in FRED with sufficient metadata yielded in total 473 instances of the present perfect with *be* and 1,062 with *have* (on the restricted data set), with 53 speakers using it with *be* at least once compared with 247 for *have*.

Features 22 and 23 are modeled in competition by means of logistic regression; the predicted odds are for present perfect realizations using *be*. As the two features are not counted over the same data set, the counts for the present perfect using *have* were scaled up to the full texts. Both models detect a significant effect for age, but with inverse directions: older speakers are more likely to use the non-standard auxiliary *be* in the lmer model, but less likely to do so in the GAM. The GAM also detects a marginally significant effect for gender such that female speakers are less likely to use *be*; the lmer model only detects a non-significant trend in the same direction. Where does this disagreement regarding age between the models come from? The overall frequency of the present perfect with *be* in most regions is very low, and the projection to the full data may well have introduced noise that the models find hard to handle. This makes the results for this feature quite suspect; for the aggregation this should not matter much, as extreme probabilities are censored. This leaves almost all counties at the bare minimum for this feature.

There is, however, a geographic distribution: the lmer county random effect has a high variance of 3.79, and the geographic smoother in the GAM is significant ($p < .001$). Map 13b illustrates the pattern: there is little variability in much of the British Isles, only Suffolk in East Anglia is markedly different and shows a much higher probability for present perfect realizations using *be*. This is somewhat surprising, considering that the feature is rated as absent in WAVE for this area. The GAM explains 75.6 percent of the deviance.

4.1.1.5. Features 24–26: modality

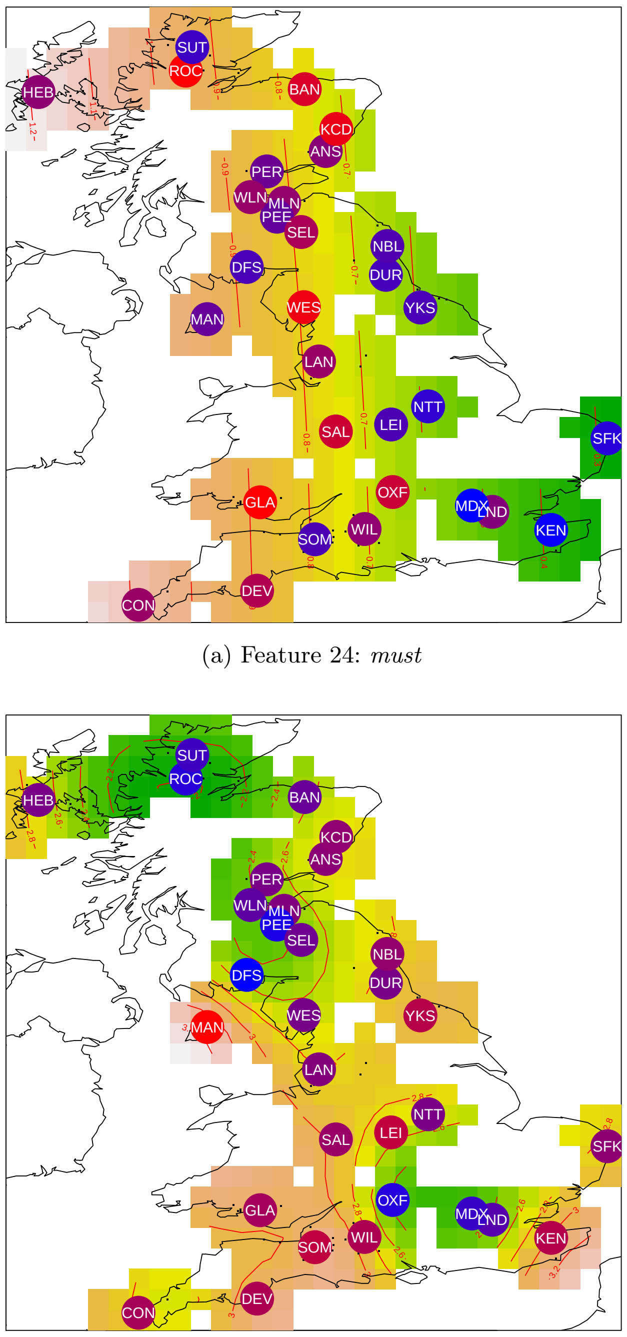
English allows several modal verbs as indicators of epistemic or deontic modality. The feature set analyzed here covers three markers, namely *must* as in (13a) to (13c), *have to* as in (13d), and *got to* as in (13e). Variation between these options in Present-day



(a) Feature 21: progressive verb forms

(b) Features 22 & 23: present perfect: *be* (predicted) vs. *have*

Map 13: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).



(a) Feature 24: *must*

(b) Feature 25: *have to*

Map 14: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

English has received considerable attention (e.g. Jankowski 2004, Close & Aarts 2010, Schulz 2012). One major result is that the frequency of *must* is on the decline, a fact which could be attributed to "a decline in forms expressing strong commitment" (Close & Aarts 2010: 178). The benefactor of this development is *have to*, with *have got to* remaining relatively stable in British English (Close & Aarts 2010: 175; Jankowski 2004: 97).

(13) a. [. . .] God has decided that he must go then, [. . .] [LAN_012]
b. Now what kind of teacher is they going to be in future years? They mustn't be as good as they was in the old times [. . .]
c. Oh I mustn't be complaining. [HEB_017]
d. We had to pick up lots of the little bits ourselves [. . .] [CON_007]
e. It had to be two buckets a day. One or t' other had got to go. [SAL_033]

Only one of these features can be linked to a feature in WAVE: F122, epistemic *mustn't* as in (13b), should in principle lead to higher usages of Feature 24, and is rated as frequent in the North of England and as neither pervasive nor rare in Scotland and the Southeast.

The text frequencies for these features were determined automatically by means of PERL scripts searching for forms of these markers. Applying this on the texts in FRED with sufficient metadata yielded in total 691 instances of *must*, 5,869 instances of *have to*, and 1,376 of *got to*. Concerning their spread, 190 speakers use *must* at least once, compared with 260 for *have to* and 215 speakers for *got to*.

4.1.1.5.1. Feature 24: marking of epistemic and deontic modality: *must*

Feature 24, *must* as a marker of epistemic or deontic modality, was analyzed using a count-based model. Both models detect the same effects of sociolinguistic predictors: female speakers use *must* more often, while older speakers have lower frequencies. There is a marginally significant trend for the interaction of gender and age in both models, such that the gender gap in *must* frequency widens with age.

There is slight geographic variability present in the data. The lmer county random effect has a somewhat low variance of 0.14, while the geographic smoother in the GAM is significant ($p < .01$). Map 14a shows that the GAM arrives at a simple east/west gradient, with *must* being more frequent in the east. The values of the lmer model and the GAM do not fit together very well here, with many of the highest county-level BLUPs, for example the northern Scottish Lowlands or Westmorland, falling into areas the GAM identifies as intermediate frequency. Clearly, the data do not support a more fine-grained pattern here. Neither is particularly harmonious with the prediction from WAVE. This is also illustrated by the low percentage of deviance that the GAM explains (8.6).

4.1.1.5.2. Feature 25: marking of epistemic and deontic modality: *have to*

Feature 25, *have to* as a marker of epistemic or deontic modality, was analyzed using a count-based model. As in Feature 24, *must*, gender has a significant effect in both models, with female speakers exhibiting greater frequency of *have to*. There is no effect of age for male speakers in both models, although the lmer model detects a significant interaction of gender and age, indicating that the gender difference increases with age. The GAM identifies the same only as a non-significant trend.

There is some geographic variability, with the lmer county random effect having a variance of 0.17, and the geographic smoother in the GAM being highly significant ($p <$ .01). Map 14b visualizes this: *have to* is most frequent in the Southwest of England and the Isle of Man, and least frequent in the Scottish Highlands and the area spanning Oxfordshire, Middlesex, and London. The Scottish Lowlands seem to show an east/west split, with frequencies of *have to* being lower in the east. The GAM explains 23.6 percent of the deviance.

4.1.1.5.3. Feature 26: marking of epistemic and deontic modality: *got to*

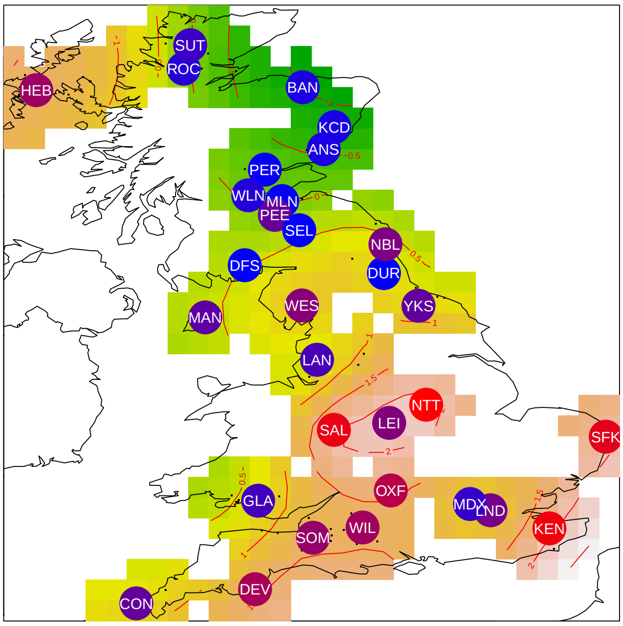
Feature 26, *got to* as a marker of epistemic or deontic modality, was analyzed using a count-based model. There are no significant effects of the sociolinguistic predictors, although the lmer model detects a marginally significant effect of age such that older speakers use *got to* less often. The GAM identifies this as a non-significant trend.

There is considerable geographic variation, with the lmer county random effect variance being 0.69 and the geographic smoother in the GAM being highly significant ($p < .001$). Map 15 depicts the distribution: there is a clear north/south split for *got to*, with the very Southeast of England and the Midlands exhibiting the highest frequencies, while the North of England shows intermediate frequencies. This feature is rare in Scotland. The GAM explains 34.7 percent of the deviance.

4.1.1.6. Features 27–30: verb morphology

4.1.1.6.1. Feature 27: *a*-prefixing on *-ing* forms

A-prefixing on *-ing* forms, as in (14), is a historical variant that is likely related to *be* + preposition + *-ing* constructions, which were available from Old English onward (Denison 1993: 388f.). Such forms are still available today in some non-standard varieties around the world (Wolfram 2008: 476f.).



Map 15: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring): *got to*. More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

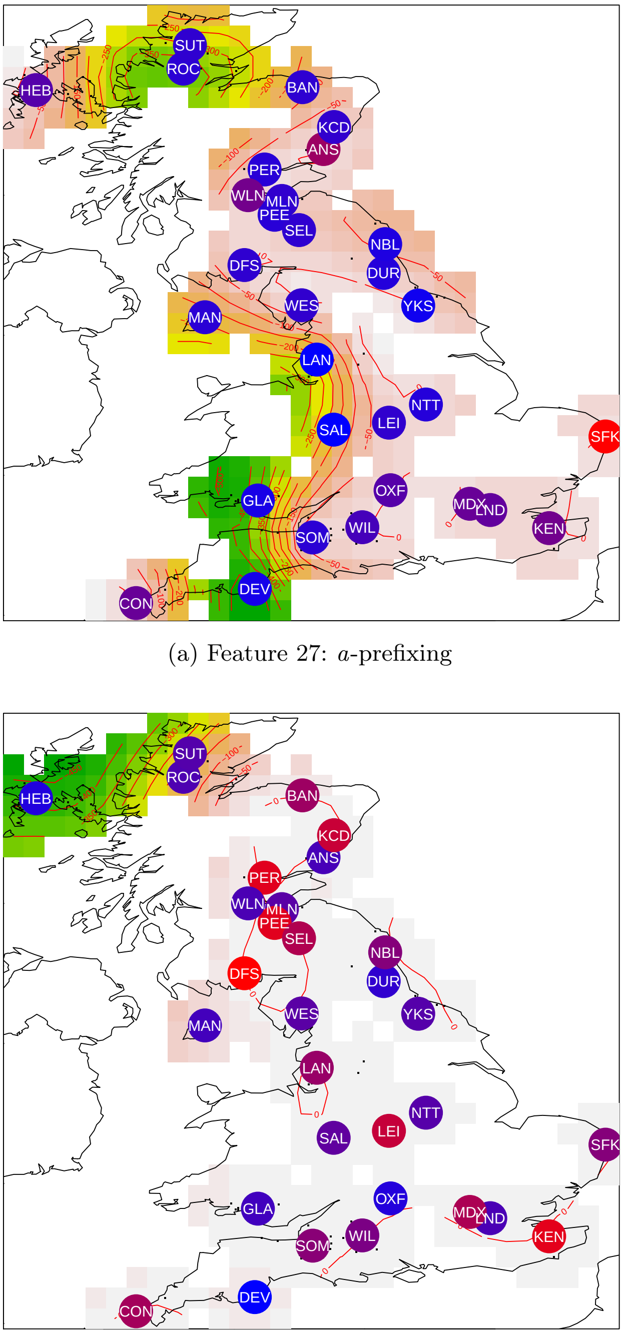
(14) [. . .] and I 'd been a-laughin' at the times when that picture was taken [. . .] [ANS_004]

The corresponding feature in WAVE is F134, also labeled "*a*-prefixing on *ing*-forms". It is rated as pervasive in East Anglia, as absent in the North and in Scotland, and as rare everywhere else.

The text frequencies for this feature were determined automatically using a PERL script collecting all instances of words ending in `ing/in'` that begin with `a-`; all forms of *a*-prefixing in the corpus contain this explicit marking. Applying this on the texts in FRED with sufficient metadata yielded in total 319 instances of *a*-prefixing on *-ing* forms, and 45 speakers used it at least once.

Feature 27 was analyzed using a count-based model. Both models agree on the effect of age: unsurprisingly, older speakers use this archaic form more often. The GAM also finds a non-significant trend such that female speakers have greater frequency of *a*-prefixing.

The lmer county random effect shows considerable variance (3.57), and the geographic smoother in the GAM is highly significant ($p < .001$). Map 16a shows that this is mostly an East Anglian feature, and is also used in other dialects in the Southeast of England and in parts of Scotland. It is rare elsewhere. With the exception of Scotland, this matches



(a) Feature 27: *a*-prefixing

(b) Feature 28: non-standard weak forms

Map 16: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

the classifications in WAVE well. The GAM explains 89.9 percent of the deviance.

4.1.1.6.2. Feature 28: non-standard weak past tense and past participle forms

Many highly frequent verbs exhibit *ablaut* in their past tense and past participle forms (e.g. *sell – sold – sold*). These verbs are called *strong verbs*. In non-standard varieties, these verbs are sometimes regularized (*sell – selled — selled*) as in (15).

(15) And once selled some cattle to some dealing fellow [...] [YKS_009]

This feature corresponds to F128 in WAVE, "levelling of past tense/past participle verb forms: regularization of irregular verb paradigms". It is considered neither pervasive nor rare in all relevant areas, with the exception of East Anglia, where it is considered frequent, and the Isle of Man and the Southeast of England, where it is considered rare.

The text frequencies for this feature were determined automatically using a two-step process: first, all forms ending in `-ed` were identified in the corpus and counted. Those that appear more than 10 times in the corpus were screened manually to determine whether they are unambiguously non-standard forms. This resulted in a list of eight verbs frequently appearing in a non-standard form; the corpus was then searched for these forms automatically using PERL scripts. Applying this on the texts in FRED with sufficient metadata yielded in total 240 instances of non-standard weak past tense and past participle forms, and 74 speakers use them at least once.

Feature 28 was analyzed using a count-based model. Both models show the same effect for the sociolinguistic predictors: older speakers use significantly more non-standard weak forms. Neither gender nor its interaction with age has an effect in either model.

There is some geographic variability for this feature. The lmer county random effect has a variance of 0.92; the geographic smoother in the GAM, however, is only marginally significant ($p < 0.09$). Map 16b shows that this feature is most frequent in the Scottish Lowlands, and does not have a clear pattern in England. The GAM values here are a bit suspect, as the model finds far too extreme values for the Hebrides. Nevertheless, the model explains 29 percent of the deviance.

4.1.1.6.3. Feature 29: non-standard past tense *done*

In some English dialects, the standard past tense of *to do*, *did*, can replaced by the past participle form *done*, as in (16).

(16) All you done is sold bootlaces in the trenches, she used to say. [LND_001]

The text frequencies for this feature were determined by automatically processing all instances of `done` to remove those that clearly cannot be non-standard past tense usages, then manually selecting valid instances from the resulting list. Applying this on the texts in FRED with sufficient metadata yielded in total 571 instances, and 127 speakers used it at least once.

Feature 29 was analyzed using a count-based model. Neither model shows any effect of the sociolinguistic predictors.

There is considerable geographic variability though: the lmer county random effect has a variance of 1.76, and the geographic smoother in the GAM is highly significant ($p < .001$). As can be seen in Map 17a, this is mostly a north/south gradient, with non-standard *done* frequent in the South of England and the Isle of Man, and rare in the North of England and in Scotland. The Midlands constitute a small transition area, as can be seen by the bunching of contour lines around Shropshire and Leicestershire. The GAM explains 37.6 percent of the deviance.

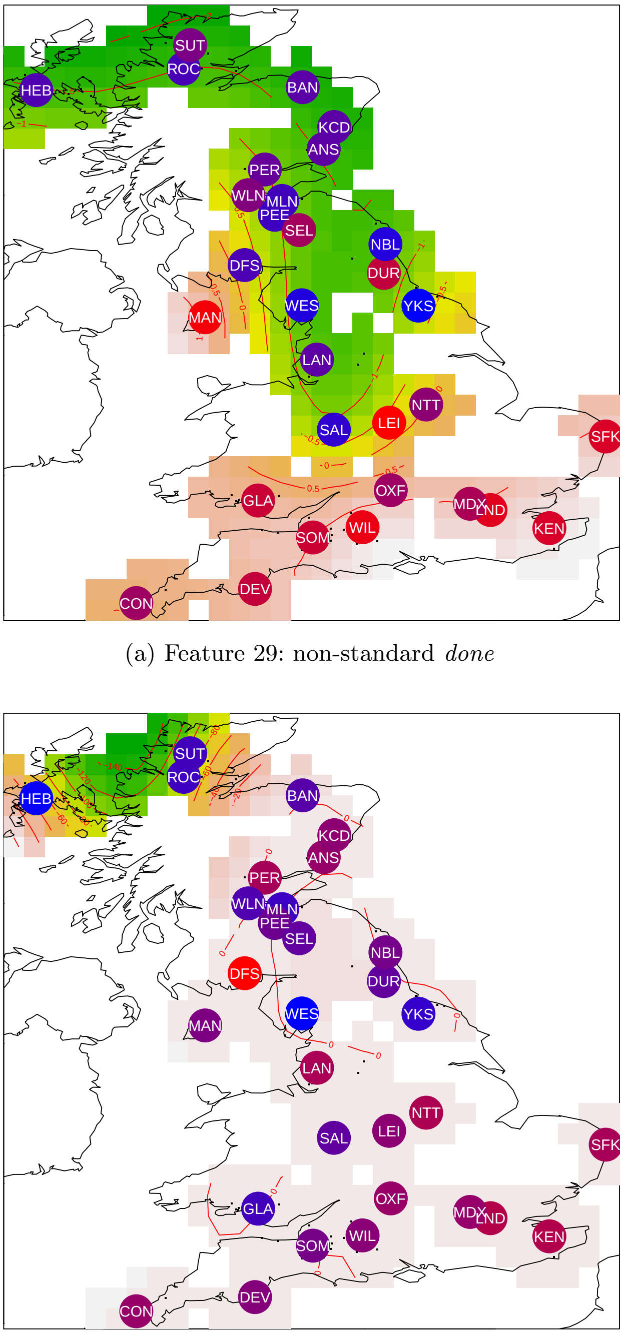
4.1.1.6.4. Feature 30: non-standard past tense *come*

The use of *come* as the past tense form instead of *came* is generally very widespread in non-standard English: Anderwald (2009: 149) attests its "enormous geographical spread". An example in British English dialects can be seen in (17).

(17) And, uh, he, he, in the end he come home on a Saturday afternoon a little bit winey [. . .] [LND_006]

The text frequencies for this feature were determined by first automatically selecting instances of `come` preceded by a third person singular pronoun or a form likely to be a name. The results were then screened manually to ensure only past tense usages. Applying this on the texts in FRED with sufficient metadata yielded in total 603 instances of non-standard past tense *come*, and 147 speakers used it at least once.

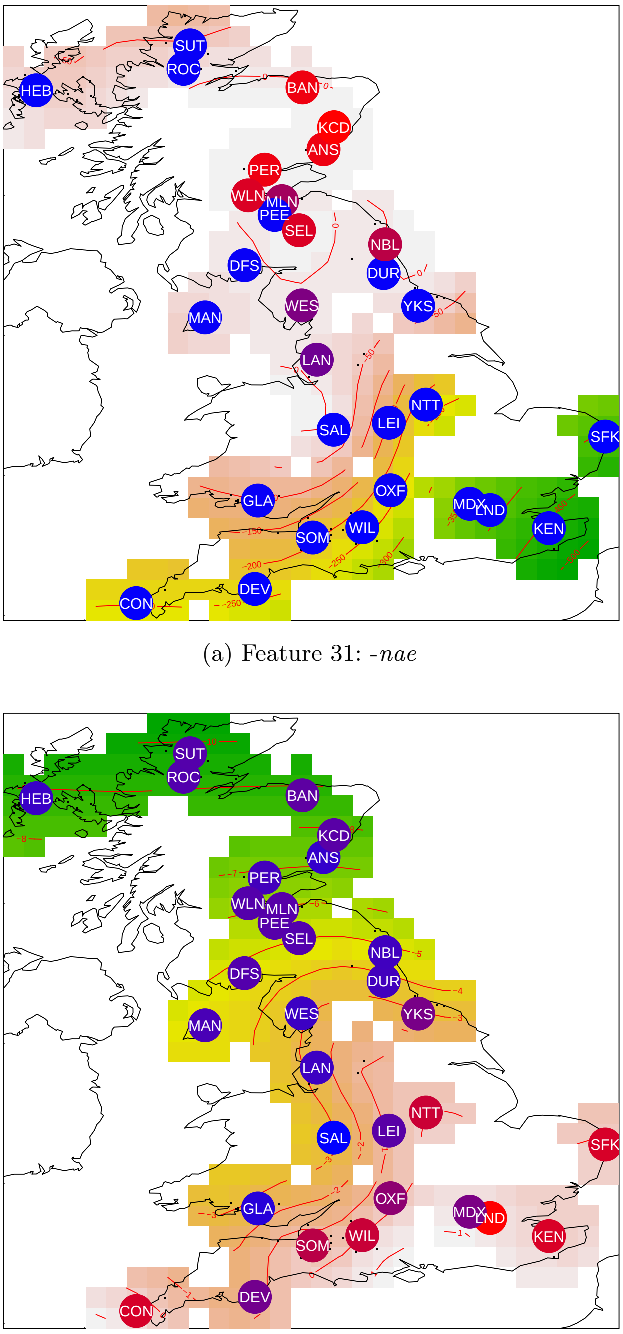
Feature 30 was analyzed using a count-based model. Both models agree that older speakers use non-standard *come* more often. The lmer model also finds a marginally significant effect of gender, such that women use it less often. The GAM agrees with this effect only numerically. There is some geographic variability: the lmer county random effect has a variance of 0.59, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 17b shows that the feature is overall a little more frequent in the south and generally less frequent the more one moves north, although there are outliers such as Dumfriesshire. The Scottish Highlands and the Hebrides show very extreme values in the GAM, making the analysis there a little suspect. Nevertheless, it explains 31.3 percent of



(a) Feature 29: non-standard *done*

(b) Feature 30: non-standard *come*

Map 17: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.



(a) Feature 31: -*nae*

(b) Feature 32: *ain't*

Map 18: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

the deviance.

4.1.1.7. Features 31–38: negation

4.1.1.7.1. Feature 31: the negative suffix *-nae*

In Scottish English, *-nae* is a negative suffix alternating with *-n't*, which can appear on all modal verbs and *do* (cf. Miller 2008: 303) and historically also appeared after full verbs (Anderwald 2003: 54). Examples can be found in (18a) and (18b). This feature is not completely restricted to Scotland, but also appears in the geographically close Northumberland, as in (18c), and very rarely in other counties[2].

(18) a. [...] but they couldnae get them to come to use their canteen [...] [WLN_006]
b. [...] we walked out on strike to get him back, we did get him back but we didnae get the wages. [WLN_006]
c. [...] I cannae remember t' schoolmaster's name [...] [NBL_003]

The text frequencies for this feature were determined automatically using a PERL script searching for word forms ending in `-nae`. Applying this on the texts in FRED with sufficient metadata yielded in total 347 instances of the negative suffix *-nae*, and 23 speakers used it at least once.

Feature 31 was analyzed using a count-based model. The models partially agree on the effect of the sociolinguistic predictors: the lmer model finds significant effects of age, indicating a decrease in the frequency for older male speakers, and of the interaction of age and gender, indicating an increase in the frequency for older female speakers. The difference between male and female speakers is only marginally significant, but quite large; female speakers use the form less, so the interactions indicate that for older speakers the gender difference is smaller. The GAM partially confirms this: the effect of age remains significant, and the interaction achieves marginal significance. The geographical distribution of *-nae* is, however, very extreme, and may cause problems for the models.

Geographic variability is, unsurprisingly, very high: the lmer county random effect has a huge variance (90.46). The geographic smoother in the GAM, on the other hand, is not significant ($p < 0.29$). As can be seen in Map 18a, the GAM contains very high adjustments in the South of England. While both models correctly identify this to be a

[2] Due to issues with model fitting, the following token from Wiltshire had to be removed from the data set:

(i) They couldnae eh keep it all going at once [...] [WIL_004] feature of the Scottish Lowlands, these high variances and smoother effects indicate that neither model can cope with the extremeness of the distribution here. Just by model fit, however, the GAM appears excellent, explaining 96.1 percent of the deviance.

4.1.1.7.2. Feature 32: the negator *ain't*

Ain't is, as Anderwald (2008: 451f.) attests, "probably the best known indicator of non-standard grammar in North America and the UK", appearing in most varieties there. In the British Isles, *ain't* can function as either the negated form of *be*, as in (19a), or of *have*, as in (19b).

(19) a. Well, draw four-thousand on account, what ain't there. [KEN_003]
b. I ain't got the time. [LND_007]

WAVE contains three features relevant to *ain't*, including the two that are attested in British English dialects: F155 and F156, which correspond to *ain't* as the negated form of *be* or *have*. Neither is attested on the Isle of Man, but F155 is considered frequent in East Anglia, neither frequent nor rare in the Southeast and the Southwest and rare everywhere else. *Ain't* as the negated form of *have* follows the same pattern, with the exception of the Southwest, where this is now rare, and Scotland, where this feature is not attested.

The text frequencies for this feature were determined automatically using a PERL script to search for the string `ain't`. Applying this on the usable parts of the data set yielded in total 185 instances of the negator *ain't*, and 60 speakers used it at least once.

Feature 32 was analyzed using a count-based model. Here, both models agree that older speakers use *ain't* significantly more often. In the lmer model, gender also emerges as significant, with female speakers using this negator much less often; in the GAM, the same effect is only a non-significant trend.

There is a clear geographic pattern in the data. The lmer county random effect has a high variance (2.54), and the geographic smoother in the GAM is highly significant ($p < .001$). Map 18b illustrates the distribution: *ain't* is primarily a feature of the Southeast of England, and also extends into the Southwest of England and the eastern Midlands. It is very rare in the North of England and in Scotland. This perfectly matches the classifications in WAVE. The GAM explains 48.1 percent of the deviance.

4.1.1.7.3. Feature 33: multiple negation

Multiple negation, also called negative concord, is the "negation of indefinite constituents in negative contexts" (Chambers 2003: 105); a very frequent feature in varieties of English around the World. Chambers (2003: 226ff.) suggests it as a vernacular primitive with a potentially innate foundation, and Trudgill (2009b: 307) argues that absence of this feature should rather be considered a peculiar feature of standard varieties. Anderwald (2005), however, notes a geographical cline in the British Isles, such that southern varieties are much more likely to use this feature than northern varieties are. Sentences (20a)-(20c) provide some examples.

(20) a. [...] course I didn't see him no more. [WIL_008]
b. Because they hadn't no parachutes love they come later all these things come later. [YKS_010]
c. We couldn't see nothing, let go an anchor, and let un go. [SOM_028]

This feature corresponds to WAVE feature F154, which is considered neither frequent nor rare in all regions except for the Isle of Man, East Anglia, and the Southwest, where it is rated as pervasive.

The text frequencies for this feature were determined in two parts: strings were identified that could conceivably constitute instances of multiple negation; these were then manually inspected to remove false positives. Strings that are, at least in FRED, always instances of multiple negation were counted directly using a PERL script. Tokens which contain more than two words between the negators were ignored. Applying this process on the texts in FRED with sufficient metadata yielded in total 1,085 instances of multiple negation, and 169 speakers used it at least once.

Feature 33 was analyzed using a count-based model. Both models find an effect of age, although it is only marginally significant in the GAM. Nevertheless, older speakers seem to use multiple negation more often.

There is considerable geographic variability: the lmer county random effect has a variance of 1.03, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 19a illustrates the distribution: multiple negation is a feature of southern England. It is much rarer in Scotland, while the North of England and the western Midlands form a transition area. Again, this is a very good match to the classifications in WAVE, with only the minor quibble that the difference between the Southeast and the Southwest of England is not readily apparent from the map. The GAM fits the data quite well and explains 46.9 percent of the deviance.

4.1.1.7.4. Features 34: contraction in negative contexts

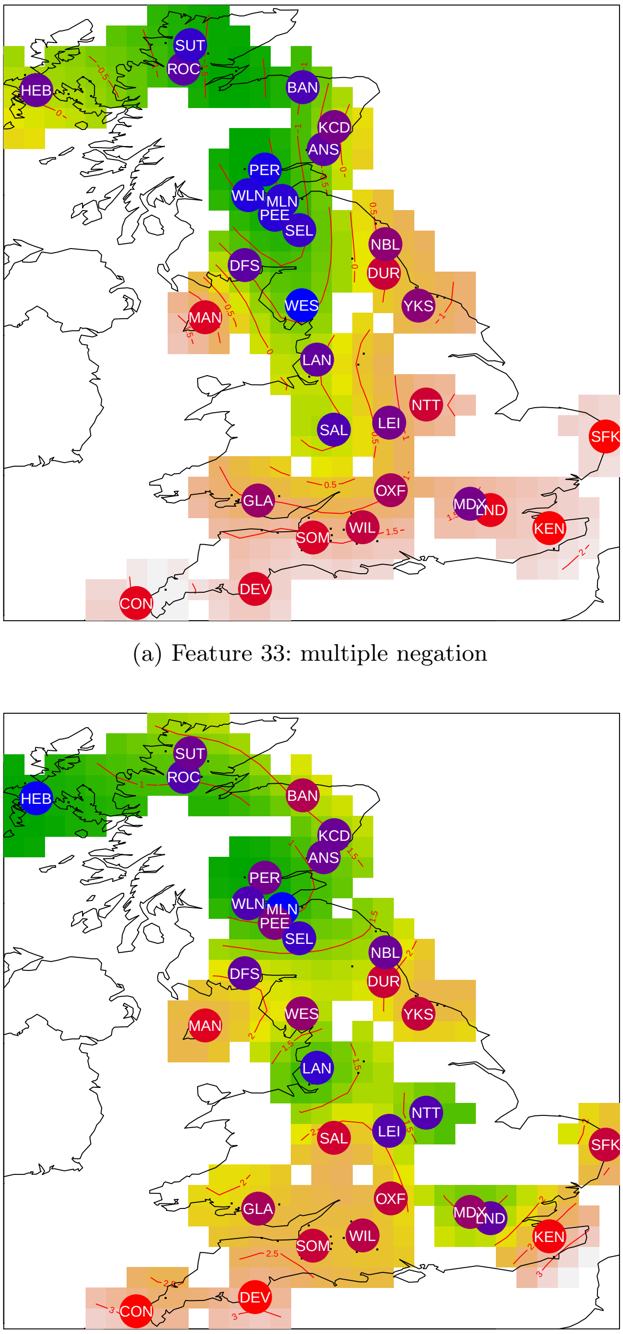
Features 34 and 35 concern themselves with contraction when both an auxiliary and a negator are involved. In these cases, the auxiliary can be realized as a full form with the negator as a contracted suffix, as in (21a) and (21b), or the auxiliary can be contracted with the negator realized as a full form, as in (21c) and (21d). Szmrecsanyi (2013: 58) provides an aggregate of the sizable literature on this topic, noting the consensus that negative contraction is more frequent in Southern English dialects, while the reverse is true for Northern English dialects. Anderwald (2002) reports that forms of present tense *be* behave differently from other verbs in that they prefer auxiliary contraction, and that there is regional differentiation but no clear geographic cline for either *be* or other verbs. She does confirm Scotland as an area of highly frequent auxiliary contraction across the board, and parts of the Midlands as high-frequency areas for auxiliary contraction with verbs other than *be*.

(21) a. But she isn't interested in that. [WIL_022]
b. But you couldn't use these now because they 're rusted, you couldn't use these. [WIL_024]
c. [...] it 's not much of a road now, but, er, they did keep what bit there was open. [SAL_027]
d. You 're not going to sing, are you, young man? [SAL_013]

The text frequencies for these features were determined automatically using a PERL script that identifies, for Feature 34, all instances of a word ending in `n't` or `nae`, and for Feature 35 all auxiliary contractions followed by `not`. Applying this on the texts in FRED with sufficient metadata yielded in total 4,625 instances of negative contraction, with 258 speakers using it at least once, and 745 instances of auxiliary contraction, which is attested for 164 speakers. The overall percentage of negative contraction is 86.1.

Features 34 and 35 were modeled in competition using logistic regression. The predicted odds are for negative contraction. Gender emerges as significant in both models, with female speakers using more negative contractions. Neither model finds an effect of age or an interaction of age and gender.

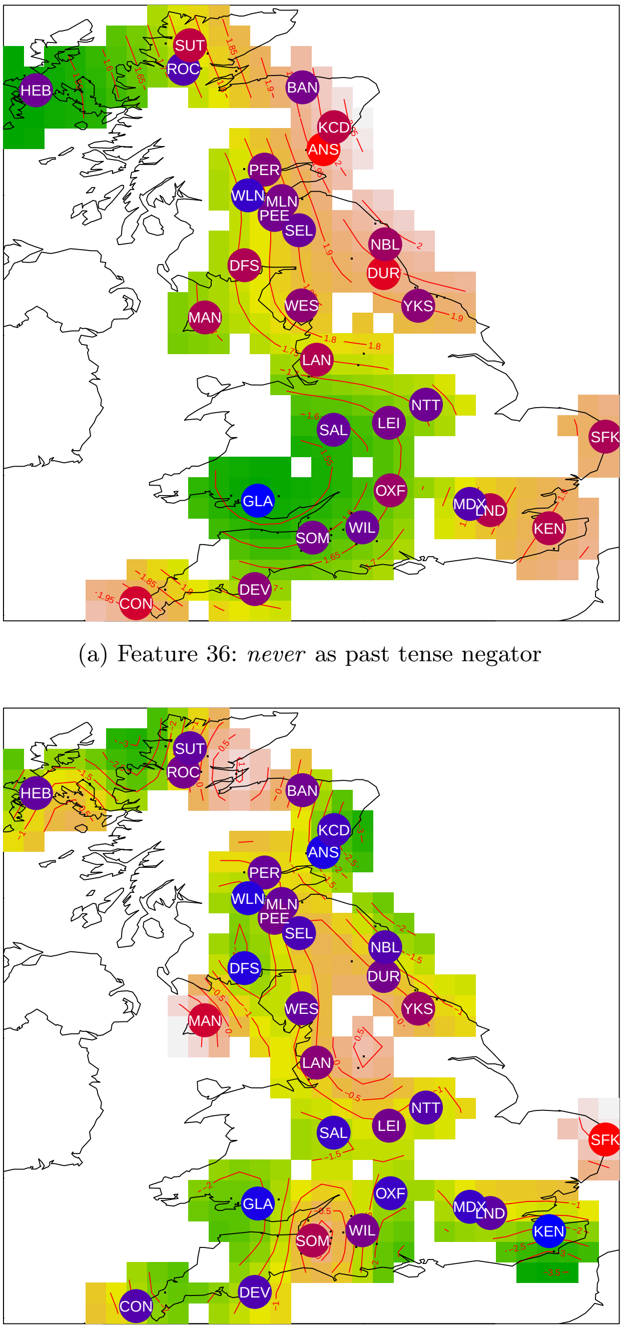
Concerning geography, the lmer county random effect has a intermediate variance of 0.5, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 19b visualizes the distribution. The very South of England emerges as strongly preferring negative contraction, a preference which decreases as one moves north. Scotland, with the exception of Banffshire, shows relatively low probabilities of negative contractions, and the same is true for Lancashire and Nottinghamshire. The transition between the North of England and Scotland is quite steep, as indicated by the bunching of contour lines near the border. This meshes quite well with the above literature on this topic. The GAM explains 33.3 percent of the deviance.



(a) Feature 33: multiple negation

(b) Features 34/35: contraction in negative contexts

Map 19: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).



(a) Feature 36: *never* as past tense negator

(b) Features 37 & 38: *was/weren't* split (predicted: *weren't*)

Map 20: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).

Section 4.1.2 will return to this feature. There, additional predictors will be used to test the robustness of this finding.

4.1.1.7.5. Feature 36: *never* as past tense negator

The use of never as a past tense negator, as in (22), is generally considered to be one of the most frequent non-standard features in British English dialects (Cheshire et al. 1995: 80) and in Englishes around the world (Kortmann & Szmrecsanyi 2004). The precise history of this feature is somewhat contentious; Cheshire (1998) sees it in direct continuation of similar forms in Middle English, while Lucas & Willis (2012) argue that it is a more recent development.

(22) So, way, he generally turned up, I 've seen him, know, being up home till eight and half past eight in the morning, but he never turned up. [DFS_001]

WAVE covers this exact feature as F159. It is considered neither frequent nor rare throughout the British Isles, with the exception of East Anglia and the Isle of Man, where it is considered pervasive.

The text frequencies for this feature were determined using a two-step process. First, all instances of `never` that could not reliably be ruled out as a past tense negator were identified automatically using a PERL script. The remaining instances were then manually screened to remove instances that were not followed by a past tense verb. Applying this on the texts in FRED with sufficient metadata yielded in total 2,023 instances of *never* as past tense negator, and 235 speakers used it at least once.

Feature 36 was analyzed using a count-based model. In the lmer model, gender and age as well as their interaction were found to be significant. Both female and older speakers use *never* more often as a past tense negator. For female speakers, the age effect is less pronounced, leading to a decrease in the gender difference for this feature with increasing age. The GAM agrees on the effect directions, but only the effect of gender is significant, while the effect of age is marginally significant and the interaction is a non-significant trend.

There is little support for an effect of geography: the lmer county random effect has a rather low variance of 0.13, and the geographic smoother in the GAM is not significant ($p < 0.25$). Nevertheless, Map 20a depicts the modeling results: there is slight evidence for a east/west gradient in the North of England and in Scotland, while in the South of England and Midlands there is a frequency valley around Somerset, Wiltshire, and Shropshire, with frequencies rising as one moves away from that area. This lack of variability again mostly matches the classifications in WAVE; even the higher frequency judgment for East

Anglia can, with some good will, be seen in the plot. Only the pervasiveness on the Isle of Man is not reflected in the models. The GAM explains a very low 8.9 percent of the deviance.

4.1.1.7.6. Features 37 and 38: *wasn't* and *weren't*

Features 37 and 38 are the first features that concern themselves with variation between *was* and *were*. Cheshire & Fox (2009) note that, across Britain, "the past BE system is reorganising towards the unambiguous expression of polarity, with *was* levelling favoured in positive polarity contexts and with parallel levelling to *weren't* in contexts of negative polarity." For areas that are further ahead in this change, we would expect fewer instances of *wasn't* and more of *weren't*.

(23) a. [...] they wasn't all that particular about that [...] [KEN_011]
b. But they weren't all at home. [YKS_002]

WAVE includes this feature as F163, "was – weren't split". It is neither frequent nor rare in all regions except for East Anglia, where it is rated as frequent, and Welsh English, where it is absent. Scottish English does not have a classification for this feature. The text frequencies for these features were determined automatically using a PERL script that searched for instances of `was` and `were` followed by `n't/nae`. Applying this on the texts in FRED with sufficient metadata yielded in total 2,077 instances of *wasn't* and 868 of *weren't*, or a percentage of 70.5 for *wasn't*. 227 speakers use the former at least once compared with 186 for the latter.

Features 37 and 38 are not necessarily in competition, but are modeled here as such, in order to approximate the *was/weren't* split that occurs in some dialects in the British Isles. The predicted odds are for *weren't*. No sociolinguistic predictors are statistically significant in either lmer model or GAM.

There is evidence for a geographic distribution: the lmer county random effect has a variance of 0.59, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 20b visualizes the distribution: *weren't* is particularly probable in East Anglia and the Isle of Man, and the probability is also higher than normal in the North of England and in parts of the English South. This matches the classifications in WAVE very well, only the higher frequency on the Isle of Man is unexpected. The GAM explains 35.5 percent of the deviance.

4.1.1.8. Features 39–45: agreement

4.1.1.8.1. Feature 39: non-standard verbal *-s*

Many English dialects extend the third person singular suffix *-s* to other persons, as in (24). Dawson (2011) provides an extensive review of the literature on this topic and the various factors that play a role in the choice of verbal *-s* in varieties of English.

(24) Like I says, the money, you had to save up for your holidays [. . .] [WIL_022]

WAVE includes this feature as F171, "invariant present tense forms due to generalization of 3rd person *-s* to all persons". For East Anglia, Scotland, and the North of England it is marked as absent, for the Southwest it is considered rare. All other regions are rated as neither frequent nor rare.

The text frequencies for this feature were determined using a two-step process: first, all tokens ending in `s` were identified and manually screened to exclude those that are not clearly verbal forms. Then, the corpus was searched for the remaining words preceded by a personal pronoun not in the third person singular. Applying this on the texts in FRED with sufficient metadata yielded in total 3,056 instances of non-standard verbal *-s*, and 189 speakers used it at least once.

Feature 39 was analyzed using a count-based model. In both models, there is a significant effect of age, such that older speakers use non-standard *-s* more often, and a significant interaction of age and gender, such that for women this difference is very close to zero. There is no significant main effect of gender, and the models do not agree on the effect direction of the trend.

Concerning geography, the lmer county random effect has a high variance of 1.03, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 21a shows the distribution: non-standard verbal *-s* is particularly infrequent in the Scottish Lowlands, and less frequent than elsewhere in the North and the central Southwest of England, as well as in East Anglia. It is particularly frequent in Nottinghamshire and in the lower Southwest of England. Once again, this matches the general pattern in WAVE, with the exception of Scotland. The GAM explains 35.6 percent of the deviance.

4.1.1.8.2. Features 40 and 41: *don't* or *doesn't* with 3rd person singular subjects

Many varieties of English generalize the form *don't* of the auxiliary *do* to the third person singular, whereas Standard English requires *doesn't*.

(25) a. He don't have any flowers on it, it 's poppies all the year round. [WIL_008]
b. He doesn't drink, [...] [LND_002]

WAVE includes this feature as F158, "invariant don't for all persons in the present tense". It is rated as frequent in East Anglia, neither frequent nor rare in Wales, the North and the Southeast of England, rare in the Southwest, and absent in Scotland.

The text frequencies for *doesn't* were determined automatically using a PERL script, by searching for the orthographic string and the corresponding Scottish form `doesnae`. The frequencies for *don't* were counted using a two-step process, first automatically selecting instances that are clearly not in the third person singular, then screening the remaining instances manually. Applying this on the texts in FRED with sufficient metadata yielded in total 115 instances of *don't* and 128 of *doesn't* with 3rd person singular subjects, or 47.3 percent *don't*. Concerning the spread, 67 speakers use the non-standard form at least once, whereas the standard form is used by 79 speakers.

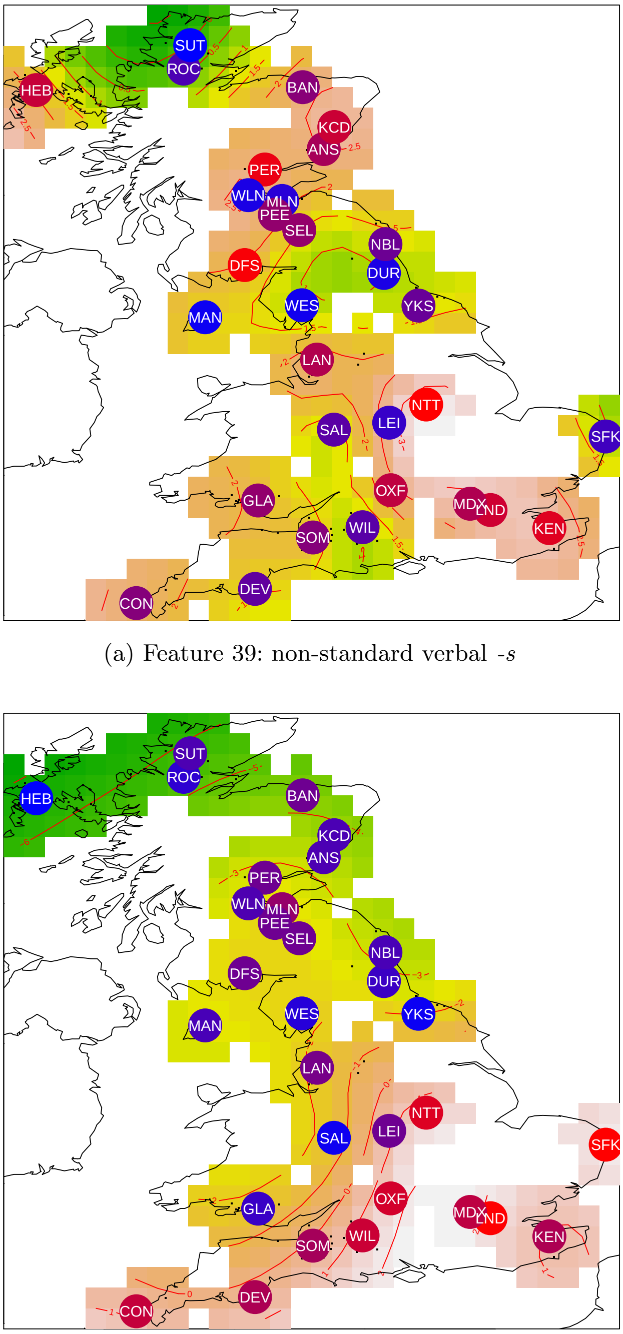
Features 40 and 41 are modeled in competition by means of logistic regression. The predicted odds are for the non-standard form *don't*. In both models, gender has an effect, such that female speakers use the non-standard form less often; neither age nor the interaction of gender and age has an effect.

There is considerable geographic variation: the lmer county random effect has a very large variance (5.26), and the geographic smoother in the GAM is highly significant ($p < .001$). Map 21b depicts the distribution. Invariant *don't* is primarily a feature of the English South, particularly of the Southeast. A rather steep transition area runs through the Midlands around Shropshire and Leicestershire. Scotland and the western part of the English North form a rather homogeneous area of low probability for invariant *don't*, and the probability further decreases toward the east and particularly the north. Again, this captures the classifications in WAVE quite well, although the frequencies are somewhat higher in the Southwest than expected. The GAM explains 56.7 percent of the deviance.

4.1.1.8.3. Feature 42: existential/presentational *there is/was* with plural subjects

Feature 42 concerns itself with usages of *there is/was* that have a plural subject, as in (26a). Standard English would require a plural auxiliary here, as in (26b).

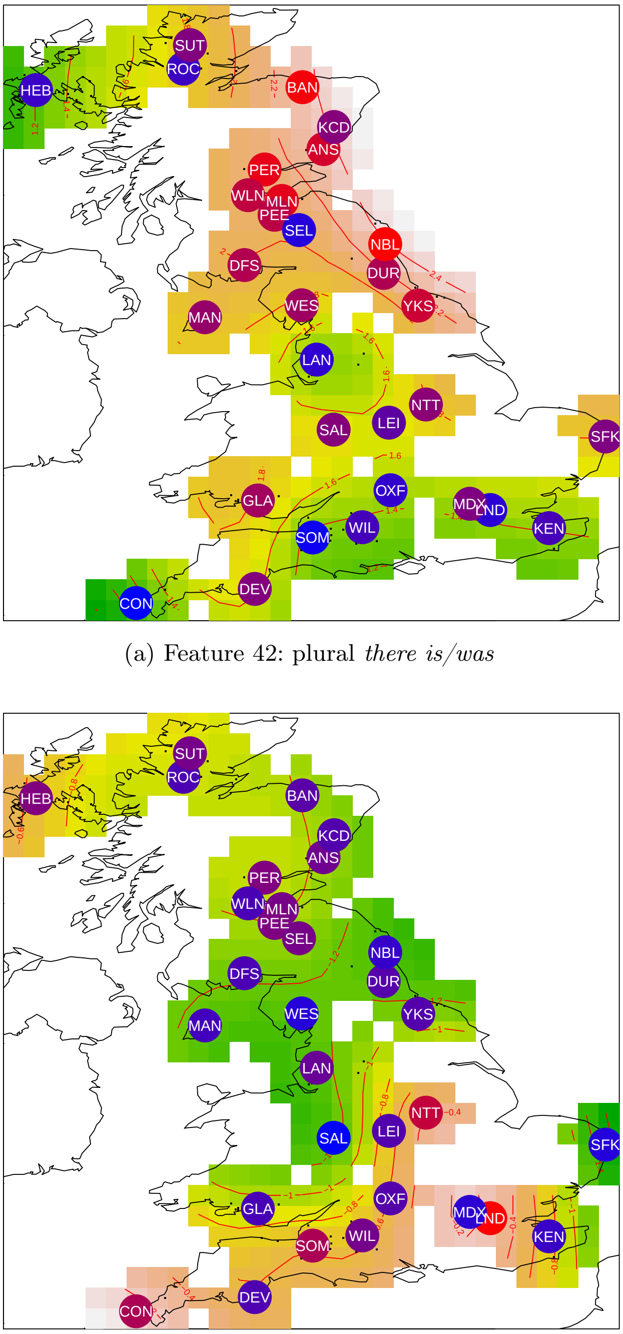
(26) a. [...] there was rockets, oh yes, we got all such as that, [...] [LEI_002]
b. [...] there were thermometers in there. [SAL_039]



(a) Feature 39: non-standard verbal *-s*

(b) Features 40 & 41: *don't* (predicted) vs. *doesn't*

Map 21: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).



(a) Feature 42: plural *there is/was*

(b) Feature 43: absence of *be* in progressive constructions

Map 22: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

This feature is included in WAVE as F172. It is considered frequent in most regions; the exceptions are the Isle of Man and Wales, where it is judged neither frequent nor rare, and the North, where it is rare.

The text frequencies for this feature were determined using a two-step process, searching for all instances of `there` followed by a singular form of *to be* that were not in turn followed by a word clearly indicating singular usage. The remaining tokens were manually screened to remove the singular subjects that were left. Applying this on the texts in FRED with sufficient metadata yielded in total 1,663 instances of existential/presentational *there is/was* with plural subjects, and 241 speakers used it at least once.

Feature 42 was analyzed using a count-based model. No sociolinguistic predictors exhibit a significant effect.

Geographically, the lmer county random effect has a rather low variance of 0.15, but the geographic smoother in the GAM is highly significant ($p < .001$). As can be seen in Map 22a, *there is/was* with plural subjects is primarily a feature of the Scottish Lowlands and parts of the North of England, especially Northumberland. As one moves south or west from there, the frequency decreases. Here, the match with the classifications in WAVE is less good: while it is rated as rare in the North of England, the map shows it to be rather frequent for most counties there. Only Lancashire clearly has lower frequencies. The GAM explains a modest 18.4 percent of the deviance.

4.1.1.8.4. Feature 43: absence of auxiliary *be* in progressive constructions

This feature covers usages of the progressive in which the auxiliary *be* is deleted, as in (27).

(27) And alright, alright, fair enough, You working? [LND_007]

This feature is included in WAVE as F174, and is absent in all regions except for the Southwest where it is unrated. Many instances of this feature are questions, and therefore fall under F228 and F229, lack of auxiliaries or inversion in *wh*-questions or main-clause *yes/no* questions (see Section 4.1.1.11.1 below). In short, for *wh*-questions, absence of the auxiliary is unattested in all regions except for the Southwest, while in *yes/no* questions it appears neither frequently nor rarely in most areas.

The text frequencies for this feature were determined using a two-step process, searching for the subject forms of personal pronouns that were followed by a word ending in `ing/in'` that could not be automatically ruled out as instances of this feature. Subsequently, the remaining tokens were screened manually to remove instances that were not progressives or where the auxiliary was present. Applying this on the texts in FRED with sufficient metadata yielded in total 126 instances of the absence of auxiliary *be* in progressive constructions, and 70 speakers used it at least once.

Feature 43 was analyzed using a count-based model. The lmer model and the GAM agree that gender has a significant effect, reducing the frequency of auxiliary deletion for women. Regarding age, the lmer model finds a significant increase for older speakers, while in the GAM this effect is only marginally significant.

The lmer county random effect has a variance of 0.47, but the geographic smoother in the GAM is not significant ($p < 0.21$). Map 22b depicts this, with frequencies higher in the central Southeast of England, and relatively low in Suffolk, the western Midlands, the North of England, and the southern Scottish Lowlands. The mismatch with WAVE is clear, even when F228 and F229 are taken into account: regions where these features are supposed to be more frequent are not clearly different from the others. Only the North of England fits the description perfectly. The GAM explains 18.8 percent of the deviance.

4.1.1.8.5. Feature 44: non-standard *was*

Features 44 and 45 continue the spectrum of variation between *was* and *were* that began with Features 37 and 38. While those features covered all uses involving negative contraction, whether standard or not, the present features concern themselves with all non-standard usages, whether negated or not. In the case of non-standard *was*, as in the examples under (28), this can be considered a case of "default singulars", another vernacular primitive according to Chambers (2003: 266).

(28) a. Well you was supposed to be, to have a batman's position [...] [LAN_020]
b. Well I thought that I would finish making that when you was here see...[WIL_024]

Both this features and the next are included in the WAVE feature set as F180, "was/were generalization". It is attested in all regions, usually neither frequent nor rare. The exceptions to this are East Anglia, where it is frequent, and the Southwest and the Isle of Man, where it is rare.

The text frequencies for this feature were determined using a two-step process, searching for all usages of `was`, including negated forms, not preceded by a word clearly indicating first or third person singular usage. The remaining list was then screened manually to remove false positives. As the number of instances of *was* is very high, the analysis was restricted to the first 1.500 words of each corpus text. Applying this on the texts in FRED with sufficient metadata yielded in total 396 instances of non-standard *was*, and 147 speakers used it at least once.

Feature 44 was analyzed using a count-based model. The lmer model detects an effect of age, such that older speakers use non-standard *was* more often; in the GAM, this effect is only a non-significant trend.

Regarding geography, the lmer county random effect has a variance of 0.3, and the geographic smoother in the GAM is significant ($p < 0.03$). Map 23a visualizes the distribution: non-standard *was* is particularly frequent in the northern Scottish Lowlands, the Southwest and Cornwall, and rare in Suffolk, the western North, the Scottish Lowlands and the Hebrides.

The GAM explains 16.3 percent of the deviance.

4.1.1.8.6. Feature 45: non-standard *were*

Another possibility is the extension of *were* into context where Standard English would require *was*, as in the examples under (29).

(29) a. She were 88 when she died. [WIL_011]
b. But he were a very nice chap. [NTT_015]

The text frequencies for this feature were determined using a two-step process. First, all instances of *were* were identified, including negated forms, unless they were preceded by a second person or plural subject pronoun. A manual screening process then removed all tokens that were not clearly first or third person singular usages. As the number of instances of *were* is very high, the analysis was restricted to the first 1.500 words of each corpus text. Applying this on the texts in FRED with sufficient metadata yielded in total 257 instances of non-standard *were*, and 77 speakers used it at least once.

Feature 45 was analyzed using a count-based model. No sociolinguistic predictors have an effect in either lmer model or GAM.

Regarding geography, the signal is more pronounced than for non-standard *was*. The lmer county random effect has a rather high variance of 3.22, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 23b illustrates this distribution. Non-standard *were* is particularly frequent in the eastern and central Midlands, especially in Leicestershire, in the eastern part of the English North, and in Somerset and Wiltshire in the Southwest of England. The GAM explains 43.8 percent of the deviance.

How do the results for Features 44 and 45 compare to the WAVE feature F180? Overall, we can see several similarities: most regions frequently use at least one non-standard variant, whether *was*, *were*, or both. The feature is marked as rare in the Southwest and on the Isle of Man, of which only the Isle of Man is obvious from the maps. Similarly, it is not apparent that East Anglia shows overall higher frequencies of *was/were* variation than the other dialect regions do.

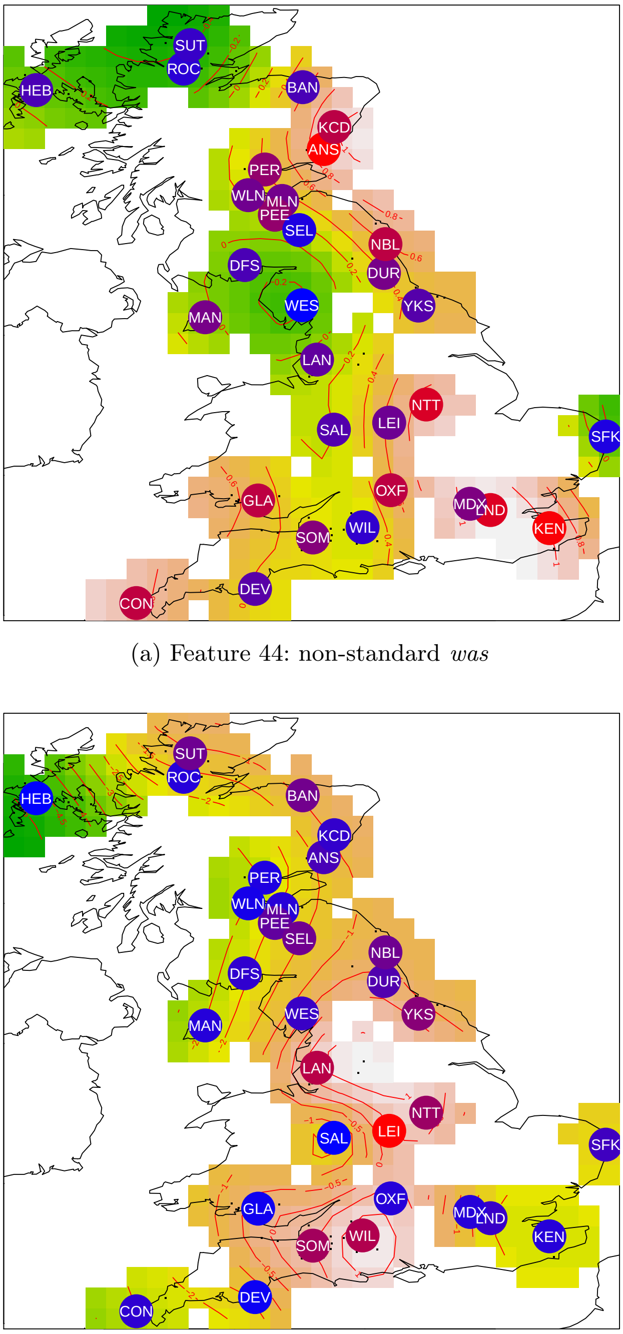
4.1.1.9. Features 46–48: relativization

Features 46–48 cover relativization markers. Feature 46 covers the standard *wh*-relativization strategy, as in (30a) and (30b), Feature 47 covers the non-standard relative marker *what* as in (30c) and (30d), and finally Feature 48 covers the relative marker *that*, as in (30e) and (30f). Herrmann (2003) provides a detailed study of relative markers in British English dialects; concerning the geographic distribution or relative markers, she finds that overall, *that* is the most frequent, but comparably rare in the Southwest and East Anglia and more frequent toward the north; *what* exhibits the inverse pattern. *Wh*-relativization is most frequent in East Anglia and the Midlands.

(30) a. [...] and there was blokes who used to come in front and blow this fire out. [SAL_025]
b. [...] to say the circumstances under which you was having to live at the time. [NTT_007]
c. Oh in them days, a pair of boots what we used to wear on the farm, they used to sell 'em in the shops four and eleven. [KEN_011]
d. I can remember a Mr Roberts what used to live down the field, [...] [SAL_037]
e. He worked for a man called Hobbs that lives down there [...] [DEV_007]
f. [...] it was something that had to be done. [HEB_035]

The relative marker *what* is included in WAVE as F190. It is considered frequent for East Anglia, neither frequent nor rare for the North, the Southeast and Wales, and rare in the Southwest and on the Isle of Man.

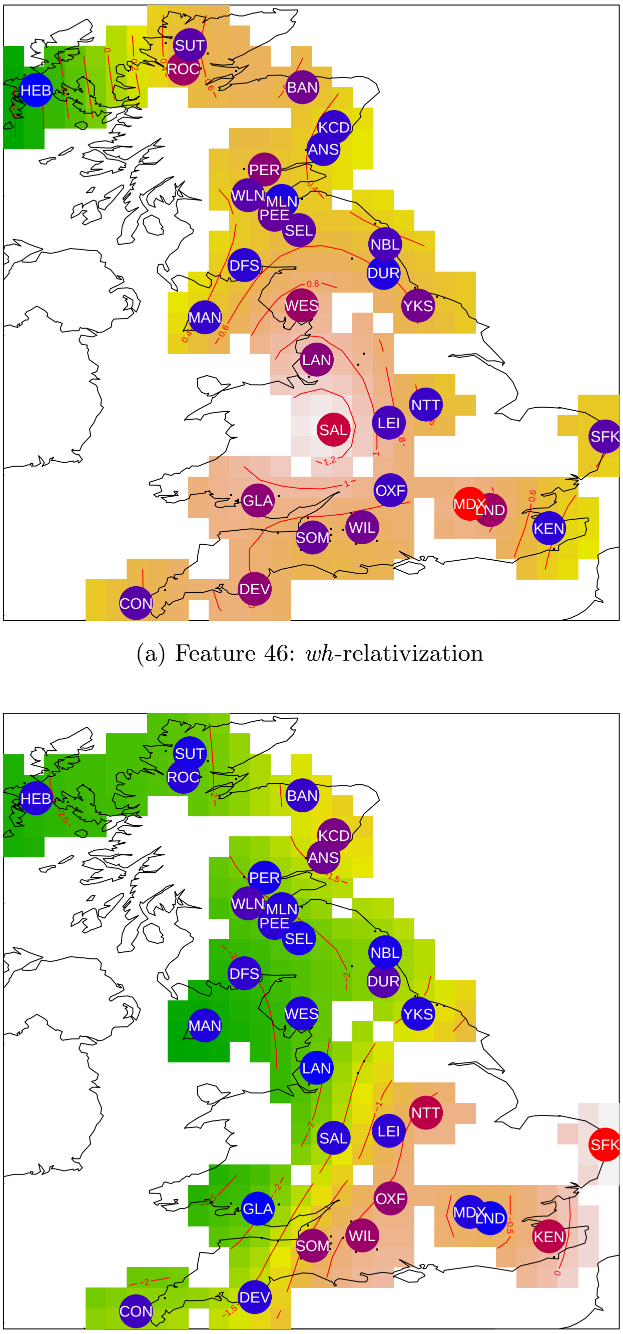
The text frequencies for these features were determined using a two-step process, first selecting all instances of *wh*-relativizers (excluding the rare *whom*), `what`, and `that`, ignoring cases that can be automatically ruled out as relative clauses. The remaining tokens were then manually inspected to remove non-relativizer usages of these tokens. As the number of instances matching a relativizer is very high, the analysis was restricted to the first 1.500 words of each corpus text. Applying this on the texts in FRED with sufficient metadata yielded in total 611 instances of *wh*-relativization by 172 speakers, 126 instances of the relative particle *what* by 66 speakers, and 615 of the relative particle *that* by 206 speakers.



(a) Feature 44: non-standard *was*

(b) Feature 45: non-standard *were*

Map 23: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.



(a) Feature 46: *wh*-relativization

(b) Feature 47: relative particle *what*

Map 24: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

4.1.1.9.1. Feature 46: *wh*-relativization

Feature 46, *wh*-relativization, was analyzed using a count-based model. Both models find a significant effect of age such that older speakers use fewer *wh*-relativizers. The interaction of gender and age, which lowers that decrease to close to zero for women, is significant in the lmer model but not in the GAM. Finally, the lmer model detects an effect of gender, with women using fewer *wh*-relativizers. This coefficient is only marginally significant in the GAM.

The lmer county random effect has a variance of 0.41, and the geographic smoother in the GAM is significant ($p < 0.03$). Map 24a illustrates these distributions. *Wh*-relativization is particularly frequent around Shropshire and rare in Scotland and the Hebrides. This somewhat matches the results from Herrmann (2003), albeit the Suffolk does not exhibit particularly high frequency. The GAM explains 13.5 percent of the deviance.

4.1.1.9.2. Feature 47: the relative particle *what*

Feature 47, the relative particle *what*, was analyzed using a count-based model. In both models, there is only a marginally significant effect of gender, such that women are less likely to use this relativizer.

Regarding geography, the lmer county random effect has a rather high variance of 0.89, and the geographic smoother in the GAM is highly significant ($p < .001$). Map 24b depicts this distribution: *what* is a feature of the English South, particularly the central Southwest, and the northern Scottish Lowlands. It is infrequent in the North of England, Cornwall, the Isle of Man, and the Hebrides. This again matches well with the pattern described in WAVE, with the exception of the northeast of the Scottish Lowlands. The results also match those by Herrmann (2003) above. The GAM explains 35.9 percent of the deviance.

4.1.1.9.3. Feature 48: the relative particle *that*

Feature 48, the relative particle *that*, was analyzed using a count-based model. No sociolinguistic predictors have a significant effect.

There is moderate geographic variability, with the lmer county random effect having a variance of 0.15, and the geographic smoother in the GAM being highly significant ($p < .001$). Map 25 shows that this is mostly a Scottish feature that is much rarer in England, with the exception of Cornwall and parts of the North and the Southeast of England. Again, the results are compatible with those presented by Herrmann (2003). The GAM explains 14.2 percent of the deviance.

4.1.1.10. Features 49–54: complementation

4.1.1.10.1. Feature 49: *as what* or *than what* in comparative clauses

Feature 49 concerns itself with either *as what* (31a) or *than what* (31b) in comparative clauses.

(31) a. [...] if they were strict to us as what they are today I would be a different man altogether! [HEB_018]
b. [...] so I says, Seek a good bit more than what you 're ever expecting to get. [PER_003]

This feature is included in WAVE as F204. For most regions it is rated as neither frequent nor rare, the exceptions being East Anglia, where it is frequent, Scottish English, where it is rare, and the Isle of Man, where it is absent.

The text frequencies for this feature were determined automatically using a PERL script that searched for all instances of the strings `as what` and `than what`. Applying this on the texts in FRED with sufficient metadata yielded in total 225 instances of *as what* or *than what*, and 103 speakers use either at least once.

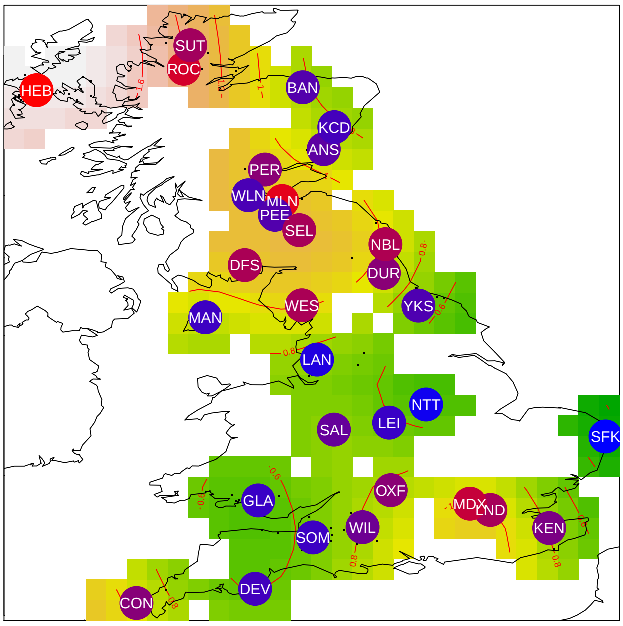
Feature 49 was analyzed using a count-based model. Both models agree that there is a gender difference, with women using this feature considerably less often.

There is only slight support for a geographic distribution of this feature. The lmer county random effect has a rather low variance of 0.18, and the geographic smoother in the GAM is not significant ($p < 0.13$). Map 26a illustrates the weak pattern, with high frequencies in Nottinghamshire and Cornwall, and low frequencies around Wales an in the northeastern parts of the Scottish Lowlands and the North of England. Nevertheless, the pattern seems compatible with WAVE, at least with regard to the lower frequency in Scotland and the general similarity of most regions. The GAM explains 14.3 percent of the deviance.

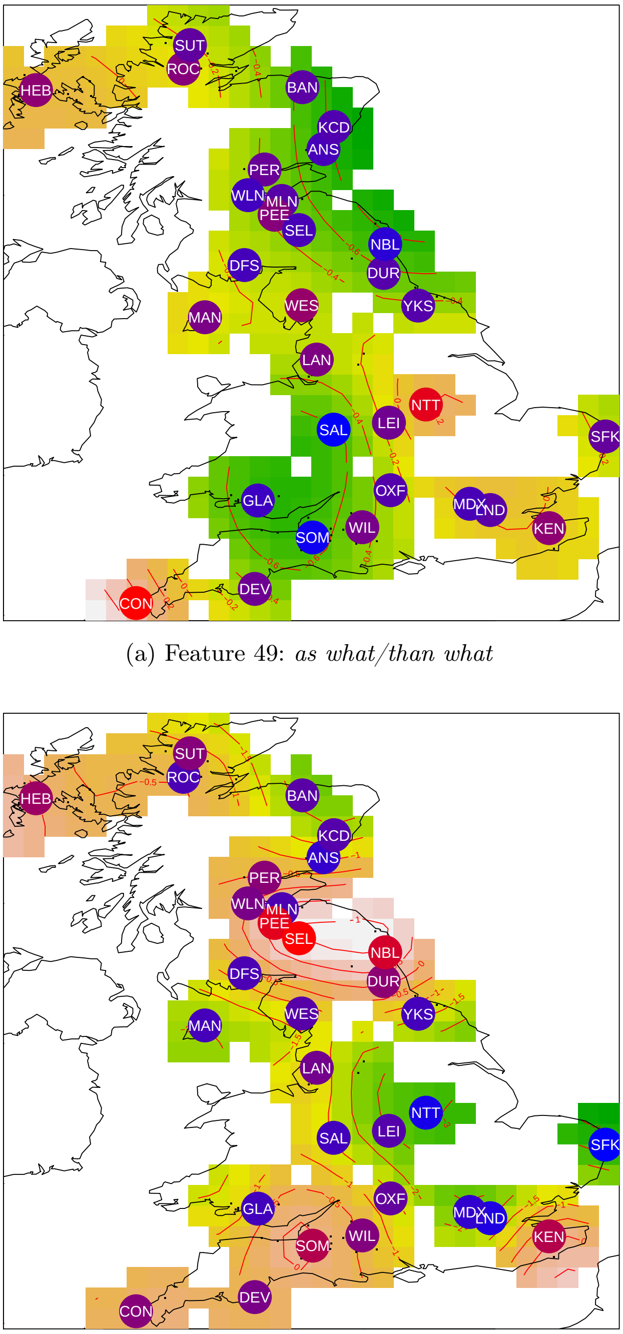
4.1.1.10.2. Feature 50: unsplit *for to*

In purposive clauses in English dialects, the infinitival marker *to* can be preceded by *for*, as is the examples under (32).

(32) a. [...] and eh, I was picked for to be on the panel, [...] [WLN_005]
b. And I used to take these loaves of bread down for to make the sausages [...] [WIL_005]



Map 25: Geographic effects in the lmer model (dot coloring) and GAM (area coloring) for the relative particle *that*. More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.



(a) Feature 49: *as what/than what*

(b) Feature 50: unsplit *for to*

Map 26: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies, more blue dots and green areas indicate lower frequencies.

WAVE contains this feature as F202, which is classified as neither frequent nor rare in Wales, the North and the Southeast of England, and rare in Scotland and the Southwest.

The text frequencies for this feature were determined automatically using a PERL script that searched for all instances of the string `for to`. Applying this on the texts in FRED with sufficient metadata yielded in total 158 instances of unsplit *for to*, and 61 speakers used it at least once.

Feature 50, was analyzed using a count-based model. Both models find a significant effect of age such that older speakers use unsplit *for to* more often than younger speakers do; in addition, the GAM finds a marginally significant trend for gender, indicating lower frequencies for female speakers.

The geographic signal in this feature is quite consistent: the lmer county random effect has a high variance of 1.7 and the geographic smoother in the GAM is significant ($p < .001$). As can be seen in Map 26b, the GAM results in high-frequency clusters around the Scottish-English border, in the Southwest and in Kent. The rest of the Southeast, the North of Scotland and the Midlands show lower frequencies. Unfortunately, this does not fit the classifications from WAVE well, particularly with regard to the Southwest. The GAM explains a respectable 39.4 percent of the deviance.

4.1.1.10.3. Features 51 and 52 : infinitival or gerundial complementation after *begin, start, continue, hate, and love*

In Standard English, some verbs allow complementation either by an infinitival verb form, as in (33a) and (33b), or by the gerund, as in (33c) and (33d).

(33) a. Well mostly women but after a few years we began to introduce a couple of men [...] [WIL_020]
b. I used to love to see all the people coming for dinner at night in this [...] [WES_006]
c. I 'm not going to start doing that. [WIL_024]
d. And we used to love going up there [...] [WES_006]

The text frequencies for this feature were determined automatically using a PERL script. The script searched for all forms of the verbs *begin, start, continue, hate* and *love*, then counted those instances followed by the infinitive marker `to` for infinitival complementation or a form ending in `ing` for gerundial complementation. Applying this process on the texts in FRED with sufficient metadata yielded in total 339 instances of infinitival complementation by 139 speakers and 500 instances of gerundial complementation by

154 speakers.

Features 51 and 52, infinitival or gerundial complementation after *begin, start, continue, hate, and love*, are modeled in competition by means of logistic regression. The predicted odds are for infinitival complementation. Both models agree on an effect of gender, such that women are more likely to use the infinitive. Age also has a significant effect in the GAM, with speakers more likely to use the infinitive the older they are; in the lmer model, this effect is marginally significant.

There is good support for an effect of geography. The lmer county random effect has a variance of 0.92, and the geographic smoother in the GAM is highly significant ($p < .001$). As Map 27a shows, the distribution is not quite clear. There are areas of higher probability for the infinitival complement in Cornwall, Shropshire, and in parts of the Scottish Lowlands, with the space in between forming valleys of higher probability for the gerund. The GAM explains 28.5 percent of the deviance.

4.1.1.10.4. Features 53 and 54: zero or *that* complementation after *think, say*, and *know*

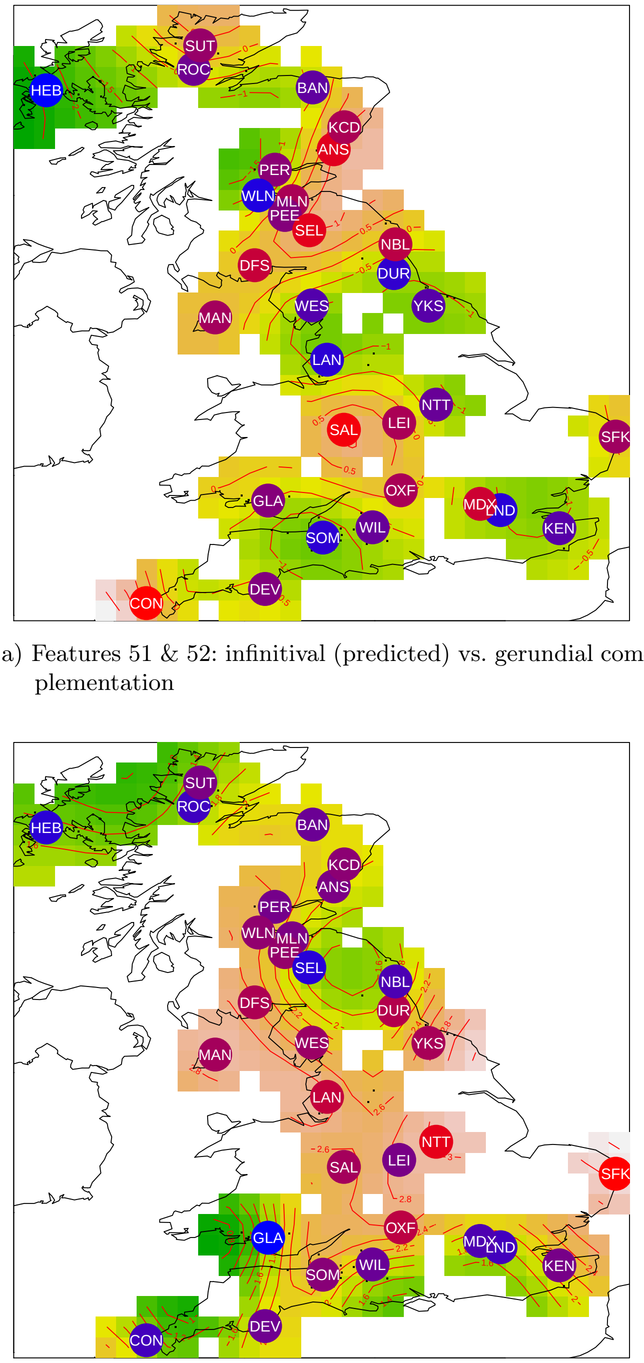
In Standard English, complement clauses can be prefixed by *that*, as in (34a), but the complementizer may also be left out, as in (34b).

(34) a. I know the coach stayed in the yard at Briery [. . .] [WES_009]
b. [. . .] I didn't know that the ducks didn't perch anywhere when I went [. . .] [YKS_011]

The text frequencies for this feature were determined automatically using a two-step process. First, all instances of forms of *think*, *say*, and *know* were identified automatically, ignoring contexts where complementation is impossible. The remaining tokens were screened manually. Applying this on the texts in FRED with sufficient metadata yielded in total 4,460 instances of zero and 421 of *that* complementation, used by 253 and 147 speakers. The overall percentage of zero complementation is 91.4.

Features 53 and 54 are modeled in competition by means of logistic regression. The predicted odds are for the zero complementation. In both models, we find no significant sociolinguistic effects; the GAM finds a marginally significant effect for gender, such that women are more likely to use explicit complementation, while the lmer model detects an interaction of gender and age, such that older women use the explicit complementation more often.

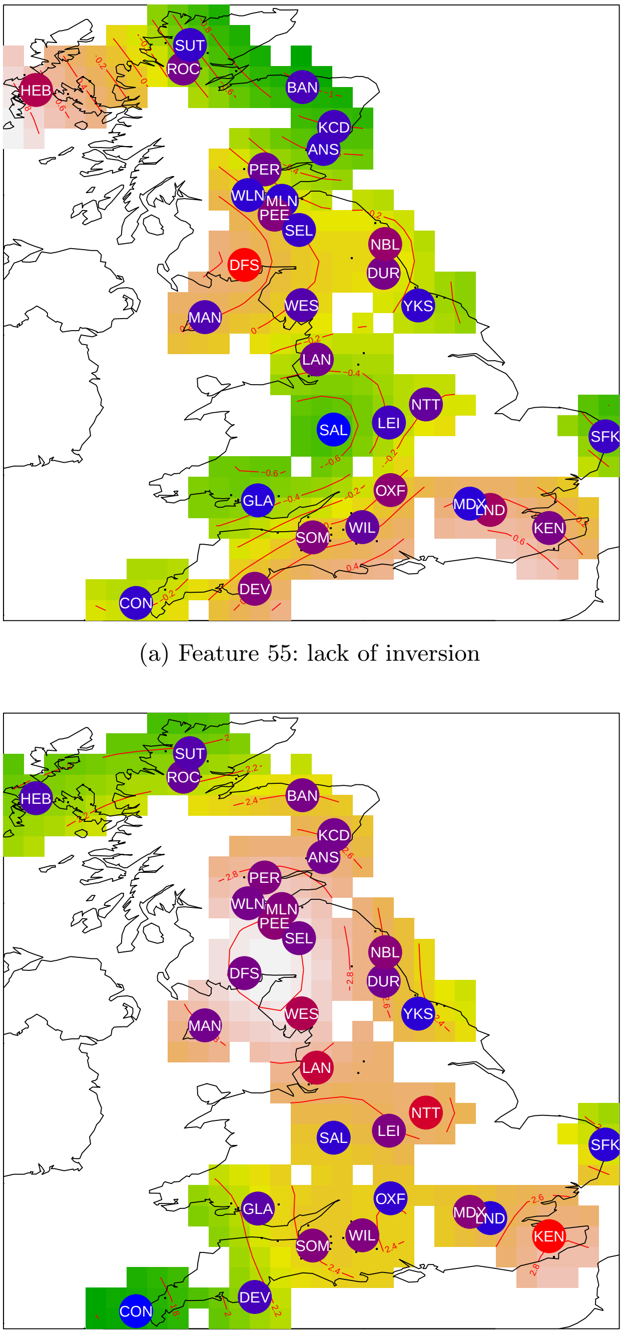
There is support for geographic variability in this feature. While the lmer county random effect has a rather low variance of 0.24, the geographic smoother in the GAM is



(a) Features 51 & 52: infinitival (predicted) vs. gerundial complementation

(b) Features 53 & 54: zero (predicted) vs. that complementation

Map 27: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher odds for the predicted realization, more blue dots and green areas indicate lower odds.



(a) Feature 55: lack of inversion

(b) Features 56 & 57: dative alternation (predicted: double object dative)

Map 28: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). More red dots and areas indicate higher frequencies (or odds for the predicted realization), more blue dots and green areas indicate lower frequencies (or odds).

highly significant ($p < .001$). Map 27b visualizes this. Zero complementation is preferred in Suffolk, the Midlands and the lower North. Further to the South, and in the Scottish Highlands and Hebrides, explicit complementation exhibits higher probabilities. The GAM explains 22.8 percent of the deviance.

4.1.1.11. Features 55–57: word order and discourse phenomena

4.1.1.11.1. Feature 55: lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes/no* questions

This feature pertains to questions where either the subject and auxiliary are not inverted, as in (35a), or where the auxiliary is completely missing, as in (35b). Kortmann & Szmrecsanyi (2004) list this as one of the most frequent non-standard features around the world.

(35) a. [...] a milk can where you used to go and get a pint of milk, you 've seen those cans? [YKS_006]
b. [...] – but, Where you put the shovel? [CON_005]

This feature is split into its components in WAVE, with F228 concerning itself with *wh*-questions only, and F229 covering the main clause *yes/no* questions. F228 is only attested in the Southwest, where it is neither frequent nor rare. F229, on the other hand, is frequent in East Anglia, neither frequent nor rare in Scotland, Wales, and the Southwest, rare on the Isle of Man, unattested in the North and not rated in the Southeast.

The text frequencies for this feature were determined using a two-step process. First, all questions were selected by searching for the question mark character `?`, ignoring irrelevant cases such as tag questions. The remaining tokens were then manually inspected to remove cases where there was inversion or that were not *wh*- or main clause *yes/no* questions. Applying this on the texts in FRED with sufficient metadata yielded in total 295 instances from 106 speakers.

Feature 55 was analyzed using a count-based model. There are no sociolinguistic effects, except for a marginally significant effect of gender in the lmer model such that female speakers tend to exhibit this feature less often than male speakers do.

There is weak support for a geographic distribution of this feature: while the lmer county random effect has a comparably high variance of 0.68, the geographic smoother in the GAM is only marginally significant ($p < 0.09$). As Map 28a illustrates, the Southern parts of England as well as Dumfriesshire have high frequencies, while Shropshire, Suffolk, and the Scottish Northwest have particularly low frequencies. Scotland and the rest of

England exhibit intermediate frequencies. Overall, this does not match the judgments in WAVE well. The Southwest, as the only region where both F228 and F229 are attested, shows higher frequency except for Cornwall, but so do Kent and London. Suffolk does not exhibit higher frequencies, but instead is one of the low-frequency areas. The GAM explains 12.1 percent of the deviance.

4.1.1.11.2. Features 56 and 57: the dative alternation following the verb *give*

English allows realization of the recipient in two major ways: either as an indirect object following the verb, the *double object* or *ditransitive dative* as in (36a), or as a prepositional phrase following the theme, the *prepositional dative* as in (36b). In non-standard varieties, the inverted order for the double object dative, as in (36c), is also available (cf. Haddican 2010). Similarly, in the case of relativized themes, there is a similar alternation, where the recipient may be marked with the preposition *to* (36e) or not (36d). Features 56 and 57 include these non-standard forms, such that datives including a preposition are included under Feature 56 and those without under Feature 57. The double object dative is the original form and allowed variation in the word order in Old English (McFadden 2002). The prepositional dative emerged in the Late Old English period, and grew in frequency during Middle English (Fischer & van der Wurff 2006). Throughout the Late Modern English period, the proportions of both variants remained quite stable (Wolk et al. 2013: Section 5). Recent research has shown that the determinants of this alternation differ in their influence between varieties of English (Bresnan & Hay 2008, Bresnan & Ford 2010, Wolk et al. 2013). As far as variability in the British Isles is concerned, Szmrecsanyi (2013: 68) gives an interpretation of the corresponding SED map, noting that the prepositional dative is especially characteristic of the Southwest and parts of the Southeast and East Anglia.

(36) a. And they gave it to them. [YKS_007]
b. aye, you gave me the money for it, you did, you gave me your money for it, I know I 've got to give mum it. [MLN_005]
c. My dad's last wage on the Gold Standard, he gave it me when I was kid, I gave it to my eldest lad twelve months or so ago. [SAL_039]
d. When mi daughter gave me a birthday party when I was eighty, lovely birthday party she gave me [. . .] [NTT_006]
e. [. . .] the beaded cape as I gave to Heritage Society at Atherton. [LAN_016]

The text frequencies for this feature were determined automatically using a two-step process. First, all instances of forms of *give* were extracted automatically, ignoring clear usages of monotransitive usages. The remaining list was inspected manually to ensure that only alternating instances remain. Applying this on the texts in FRED with sufficient metadata yielded in total 130 instances of the prepositional and 1,410 of the double object dative, or a percentage of 8.4 prepositional realizations of the dative. 75 speakers use the prepositional dative, and 222 the double object dative.

Features 56 and 57 are modeled in competition by means of logistic regression. The predicted odds are for the double object dative. Again, both models agree on the effects of sociolinguistic predictors, with age, gender and their interaction emerging as significant. Both women and older males tend to use the prepositional dative more often; the increase with age is lower for women, indicating a reduction of the gender difference.

There is very little geographic variation: the lmer county random effect has a low variance of 0.06, and the geographic smoother in the GAM is not significant ($p < 0.44$). As can be seen in Map 28b, the prepositional dative seems to be less probable in Suffolk, the very Southwest, the Hebrides, and the Scottish Lowlands, and it is more frequent in Kent, Nottinghamshire, the southern Scottish Lowlands and the western parts of the North of England. Despite the low reliability, the results seem compatible with the SED data. The GAM explains a rather low 10.3 percent of the deviance.

4.1.2. On the effect of additional predictors

The regression models in the previous section contained only few predictors: geographic location, speaker gender and age. Compared to the richness that is characteristic of sociolinguistic and probabilistic studies of language variation, these models seem overly simplistic. Furthermore, some features, such as Feature 5, the personal pronoun *us*, do not measure the phenomenon (the non-standard usages of *us*) directly, but a super-set of the possible instances of that phenomenon. This approach relies on the intuition that, everything else being equal, frequency differences for that phenomenon will percolate upward to the total frequencies, and that using a large corpus ensures that everything else is sufficiently equal. As we have seen so far, the results of this approach largely overlap with the results from more traditional investigation, and therefore seem justified.

In this section, I present a more detailed investigation of one feature, the choice between negative and auxiliary contraction. This allows for an investigation of how the simple model fares in comparison to a more sophisticated one, and whether simple modeling is an improvement over using normalized values. I will begin by defining the variable context in more detail, and by adding some linguistic predictors that have emerged as particularly relevant in the literature. The major point of reference is Tagliamonte & Smith (2002), who compiled a corpus of sociolinguistic interviews at eight locations in

England, Scotland, and Northern Ireland. Their central finding is that there are dialectal differences in contraction choice, but the pattern is not as clear as often claimed: neither the north/south nor the England/Scotland distinction are adequate descriptions.

4.1.2.1. The variable context

Szmrecsanyi's feature extraction process here is quite simple: the number of relevant auxiliary contractions is determined by searching for all contracted auxiliaries followed by `not`, while the number of negative contractions is determined by counting all forms ending in a contracted negator.

Tagliamonte & Smith (2002) provide a much more detailed variable context. I apply their definition, as appropriate, to refine the counts provided by Szmrecsanyi's method.

Verbs

Three auxiliaries can undergo both auxiliary and negative contraction: *be, have,* and *will* (Tagliamonte & Smith 2002: 257f.)[3]. Of these, *be* is considered to vary freely, while the others strongly prefer negative contraction. The new data set only contains forms of these verbs, with the exception of *ain't* which can function as the negative contraction of both *be* and *have. Be* only allows negative contraction outside the present tense, as there are no contracted forms of *was* or *were.* Therefore, all *wasn't/weren't* tokens were excluded from the data set, as well as the non-standard form *warn't.* Furthermore, Tagliamonte & Smith (2002: 257f.) report that first person subjects of present tense *be* behave differently. As Quirk et al. (1985: 129) note, there is no "completely natural" form of negative contracted *I am not*, and the common non-standard forms were very rare in Tagliamonte & Smith's data, with no observations for *amn't*, 7 for *aren't*, and 12 for *ain't*, a form that only appears in two of their sites. Therefore, their variable context excludes first person *be.* In the FRED data set, however, this is not true. There are 15 observations of *ain't* with first person subjects, spread over 7 of the counties, and 28 of *aren't* in 4 counties. Compared to the 272 instances of *I'm not* and *I's not*, over 13 percent of the first person present tense forms of *be* involve auxiliary contraction. Therefore, these cases were not excluded from the analysis.

Tagliamonte & Smith (2002: 257f.) do include *would*, but note that this auxiliary is commonly considered to not alternate, and they only find a single case where *would* contracts to *'d not.* In FRED, there are 9 instances of this, as in the examples in (37).

[3] *shall not* and *should not* can, in principle, also be contracted to *'ll not/shan't* or *'d not/shouldn't.* There were no relevant cases where *'d* not clearly means *should.* Contraction of *shall* can be difficult to distinguish from *will*, but is generally considered rare today (Quirk et al. 1985: 122).

Compared to over 1000 instances of *wouldn't*, their frequency is so low that they do not convey useful frequency information, and both are excluded from the analysis.

(37) a. Yeah yeah we 'd not be here any more. [ANS_004]
b. It 's, it 's alright he 'd not bother [LAN_005]

Interrogatives and tag questions

Tag questions and interrogatives often involve negative contraction, as in the examples under (38). In neither case is the alternation with auxiliary contraction available, therefore these constructions should be excluded from the analysis, as suggested by Tagliamonte & Smith (2002: 263f.).

(38) a. Over east Trevegian, isn't he? [CON_006]
b. Isn't it awful how things disappear ... [WES_008]

An inspection of the data showed that the vast majority of tag questions and interrogatives in this data set involve pronouns, more specifically, they are immediately followed by a pronominal form. This led to the following heuristic: a token was marked as potentially being a tag question if that token was immediately followed by a personal pronoun, *there* or *here*. This list also contained those non-standard pronoun forms that appeared after a contraction in the data, such as *'t* for *it* or *'e/ee* . These tokens were then removed from the data set if they involved a negative contraction.

To evaluate this heuristic a sample of 50 tokens that were included and a sample of 50 auxiliary contraction tokens that remained were drawn from the data set. The sample of remaining tokens contained no interrogatives or tag questions, suggesting that the heuristic leads to few false inclusions. Among the excluded sample, two instances of false exclusions were found, where a clause boundary intervened between the negative contraction and the pronoun, as in (39a). Furthermore, this heuristic led to the exclusion of two instances of disfluencies, as in (39b). Note that the second instance of *won't* is included as another token in the data set. The exclusion of these tokens therefore improves the data, as these two instances of *won't* are not two independent tokens. There is, however, the problem that this might add a bias against negative contractions. Therefore the auxiliary contractions that were followed and preceded by a pronominal form were checked, and no comparable instances were found, suggesting that this particular issue does not usually affect auxiliary contractions. We therefore end up with two true false exclusions out of 50, a rather low rate.

(39) a. Well no with the seniors we hadn't we used to turn out of a Saturday and that was it. [WES_018]
b. no no no he won't he won't make them. [WES_003]

Auxiliary deletion and null subjects

Tagliamonte & Smith (2002: 263) also exclude sentences with null subjects, as in (40a), and those where the auxiliary is deleted (40b). As the process employed here searches for auxiliaries, deleted ones were never included in the data set. There was no complete screening for null subjects, although cases in which the auxiliary was clearly sentence-initial were removed. This affected 4 interrogatives and only 1 case of a null subject. This suggests that overall, null subjects are not particularly common in this data set.

(40) a. Ain't none. [NTT_013]
b. Well, there Ø no many able to dance. (Tagliamonte & Smith 2002: (17a))

Predictors

This section describes the predictors that were used to model this alternation. Two factors that Tagliamonte & Smith (2002) code for, complement type as well as usage as auxiliary or copula in the case of *be*, are not included here. Both require extensive manual coding, and were found to have no significant effect in their data.

Speaker Age, Gender, and Location

These predictors are as described in Section 3.1.1. Speaker age is centered around the mean age to make the default values more easily interpretable and reduce potential problems with multi-collinearity. Speaker gender is a binary predictor, with "male" as the default level. Location is operationalized using the county labels for the LMER models and the interview location's coordinates for the GAMs.

Two Scottish locations with relatively little data were removed from the analysis due to the more restricted variable context: Banffshire and Kincardineshire. Of the original tokens from these locations, not a single one allowed alternation per the definition above.

Verb

As discussed above, three different auxiliaries are included: *be*, *have*, and *will*. In most cases, the form of the auxiliary can be directly matched to an auxiliary type. The exceptions are `'s`, which can stand for *is* or *has*, `'d` which can stand for *had* or *would*, and `ain't`, which can function as the negated form of *be*, *have*, and *do*. These were manually disambiguated.

In total, the data set contains 1016 tokens of *be* (650 auxiliary contractions), 877 tokens of *have* (23 auxiliary contractions), and 233 tokens of *will* (40 auxiliary contractions).

Preceding context

Tagliamonte & Smith (2002: 261) include two predictors concerning the preceding context: first, whether the subject type is nominal or pronominal, with pronominal forms appearing more often with auxiliary contraction. Second, they found that the phonological environment matters; when preceded by a vowel, auxiliary contraction becomes more probable.

Both predictors are included in this analysis. The subject type is approximated by looking at the word immediately preceding the auxiliary. In almost 90 percent of cases this identified either a personal pronoun, a demonstrative pronoun, or *here* or *there*, which Tagliamonte & Smith (2002: 261) include as pronouns. These cases were marked as pronominal subjects, and all others were marked as non-pronominal. This heuristic is not necessarily accurate; nevertheless, a sample showed that the heuristic performs quite well, with only a single case where an immediately preceding pronoun is not the subject. Finally, first person singular pronouns were labeled as a separate category; this can capture some of the variability resulting from the lack of a widespread first person negative contracted version of *be*.

Determining the vocality of the previous sound is difficult from written material. It can, however, be approximated using computer-based pronunciation dictionaries. The dictionary chosen here is the UNISYN dictionary[4], provided by the University of Edinburgh. While UNISYN is available for several British and international dialects, not all areas in this sample are covered, and therefore the RP version of the dictionary was used. One problem concerns rhoticity: while most English dialects are non-rhotic, a few of them as well as the Scottish dialects are. Furthermore, the areas are not stable, as non-rhoticity is spreading eastward in England (Chambers & Trudgill 1998: 95). Upton (2008: 280) lists Scotland and the Southwest of England as the regions in this data set that clearly still exhibit rhoticity; in these areas, word-final /r/ was kept. For the other regions, three options were explored: keeping every /r/, no /r/, or those where the following word starts with a vowel to simulate linking /r/. The most conservative choice here is to keep none, and this is the default choice for the analyses presented below.

[4] Available online at `www.cstr.ed.ac.uk/projects/unisyn/`.

	Coefficient	SE	Z	p
(Intercept)	−2.27	0.38	−6.0	**<.0001**
gender: female	0.16	0.21	0.7	>0.5
age (mean 0)	0.02	0.01	2.2	**<.05**
previous word: non-pronominal	3.55	0.41	8.7	**<.0001**
previous word: pronominal	1.04	0.19	5.6	**<.0001**
verb: have	5.61	0.27	20.6	**<.0001**
verb: will	3.32	0.24	13.9	**<.0001**
preceded by vowel	−0.35	0.19	−1.9	<0.1
gender/age interaction	−0.02	0.02	−1.5	>0.1

Table 4.1.: Contraction: coefficients of the lmer model. Predicted odds are for negative contraction, i.e. positive coefficients indicate increased probability of negative contraction. Significant predictors highlighted in bold.

4.1.2.2. Results

Let me begin by recapitulating the results obtained from the model with no language-internal predictors. Both lmer model and GAM found that the only significant sociolinguistic predictor was age. Model quality was acceptable, with the GAM explaining 33.3 percent of the deviance, and a geographical distribution was present: the lmer county random effect had a variance of 0.5, and the GAM smoother was significant. This distribution was such that auxiliary contraction was most frequent in England, with the very Southeast and Southwest having particularly high rates. Scotland, and the Lancashire-London axis emerged as hot spots for negative contraction.

Table 4.1 shows the result of the lmer model on this data set. The variability of the county random effect is considerably higher than before, at a variance of 2.0 instead of 0.5. This suggests that, taking the additional predictors into account, the geographic differences increase. The model quality is good: 87.9 percent of tokens are predicted correctly. This is a considerable improvement over the baseline of 65.1 percent, which results from always predicting the most frequent realization in the data set, negative contraction. The C value of this model is satisfactory as well: at 0.87 it is comfortably over the customary threshold of 0.8, suggesting that the model is useful in predicting the response (Baayen 2008: 204).

Table 4.2 displays the results of the GAM model. Again, the model is quite good, with virtually the same C score of 0.87, and a slight improvement in predictive accuracy to 88 percent (with the same baseline of 65 percent). The geographic effect remains significant. The model explains a considerable amount of the deviance at 56.6 percent, an improvement

	Coefficient	SE	Z	*p*
(Intercept)	-1.89	0.24	-7.82	**<.0001**
gender: female	0.09	0.21	0.43	>0.5
age (mean 0)	0.03	0.01	2.86	**<.01**
previous word: non-pronominal	3.54	0.4	8.82	**<.0001**
previous word: pronominal	1.07	0.18	5.78	**<.0001**
verb: have	5.53	0.27	20.8	**<.0001**
verb: will	3.24	0.23	13.81	**<.0001**
preceded by vowel	-0.38	0.19	-2.04	**<.05**
gender/age interaction	-0.02	0.02	-1.42	>0.1

Table 4.2.: Contraction: coefficients of the GAM. Predicted odds are for negative contraction, i.e. positive coefficients indicate increased probability of negative contraction. Significant predictors highlighted in bold.

over the previous GAM at 33.3 percent. How much of this can be attributed to geography versus the other factors? To test this, a model containing no geographic information was built. This model fares considerably worse: it only predicts 82.5 percent of tokens correctly, has a C value of 0.81 and explains only 44.4 percent of the deviance. Therefore, the geographic distribution is not only statistically significant, but also meaningful in practice.

Let us now turn to the effects found in the model. In general, both models closely agree on both effect directions and sizes. Concerning the sociolinguistic predictors, there is a notable difference between the models presented here and those discussed in Section 4.1.1.7.4. Previously, we found a significant effect of speaker gender, but not age. This pattern is reversed here: there is no significant effect involving gender, but one involving age. This does not necessarily mean that there is no gender difference, especially as the coefficients continue to point in the same direction. However, it does suggest that part of the previously observed effect of gender was confounded with other factors, such as the inclusion of tag questions. The effect of age found here is such that older speakers are more likely to use negative contraction.

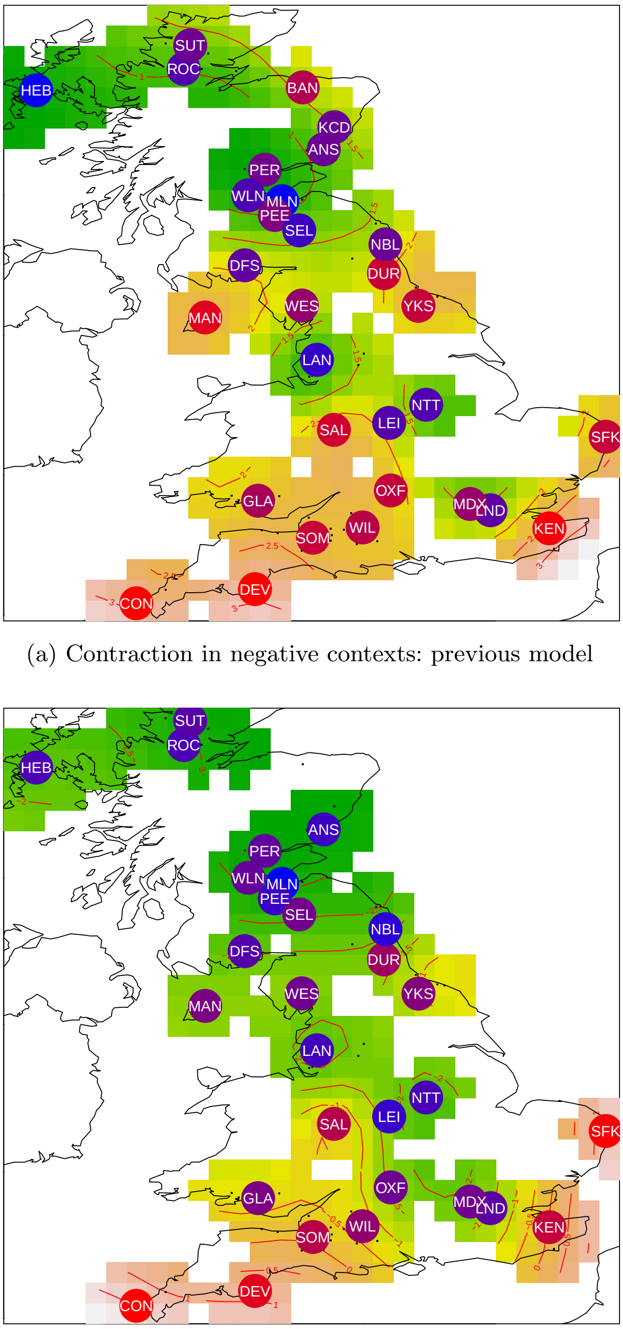
Concerning the verb, we find that both *have* and *will* are more likely to appear with negative contraction than *be*, and that this tendency is strongest for *have*. This result is in lockstep with those reported by Tagliamonte & Smith (2002) and Anderwald (2002: 80).

Concerning the preceding context, Tagliamonte & Smith (2002) found, through cross-tabulation, that the effect of pronominality is really one of the phonological environment, and they therefore include only the vocality of the previous sound in their VARBRUL analysis. In contrast, Bridge (2006) reports that in Derby in the Northern Midlands, pronominality has a stronger effect on alternation choice and that the phonological environment does not have a significant impact. On the present data set, both factors matter: for pronominality of the preceding word, the first person pronoun favors auxiliary contraction most strongly, followed by other pronouns and finally non-pronominal constituents, which overwhelmingly favor negative contraction. Concerning the preceding sound, the lmer model finds a marginally significant effect, which is significant in the GAM. It suggests that, as in Tagliamonte & Smith (2002), a preceding vowel leads to more auxiliary contraction. This effect is, overall, relatively small, which may be an effect of the conservative coding for rhoticity, as discussed above. Including linking /r/, or counting more dialect areas as rhotic, would increase the effect of this predictor, yet the fact that pronominality has a large influence on the result does not change. The fact that the phonological environment does emerge as significant does suggest that both factors have an effect independently from one another.

I now turn to the geographic distribution of contraction choice. The result of the basic models can be found in Map 19b, reprinted here for convenience as Map 29a. It was found that, in general, the Southern dialects employ more negative contraction, a tendency that decreases as one moves north. Scotland, as well as the Lancashire–London axis, exhibited particularly high probabilities for auxiliary contraction. This finding largely matches the consensus in much of the literature, such as Hughes & Trudgill (1979: 20f.)[5].

The results of the new models are visually very similar to the previous ones, confirming this overall pattern and strengthening it. While the distribution of high and low values are similar, the newer model has larger differences, as evidenced by the greater number of contour lines. Removing non-alternating instances of negative contraction, such as tag questions, and removing variation that can be explained by other factors makes the overall pattern clearer. This is particularly obvious in the case of the northern Scottish Lowlands. On the original data set, a particularly high value for negative contraction was found in Banffshire, leading to a northeast/southwest cline in Scotland. This high value resulted from the fact that the data for Banffshire contained no instances where the alternation was truly possible. Removal of these cases led to a clearer clear north/south distinction in a wide area around the Scottish border.

[5] Hughes & Trudgill, however, limit their claim to words other than *be*.



(a) Contraction in negative contexts: previous model

(b) Contraction in negative contexts: detailed model

Map 29: Geographic effects in the lmer models (dot coloring) and GAMs (area coloring). Lighter and more red colors indicate higher odds for negative contraction.

4.1.2.3. Discussion

As we have seen the more elaborate model confirms the geographic pattern found in the simple model. This raises two questions: first, what do the results tell us about the variation between auxiliary and negative contraction in British English? And second, what are the implications of this for the application of simple modeling to dialectometric analysis?

To answer the first question: one major result is that three predictors that were found to have a major influence in previous studies were confirmed here. The type of verb was found to have a large influence (Tagliamonte & Smith 2002, Anderwald 2002), as did the preceding context in terms of pronominality (Bridge 2006) and, to a lesser degree, phonology (Tagliamonte & Smith 2002). The locus of variability resides mostly in present tense *be*, with *will* showing considerably less variation and *have* being almost categorically associated with negative contraction. This is again in line with previous research on the topic. However, Anderwald's finding that throughout Britain the percentage of auxiliary contraction for present tense *be* lies above 80 (2002: 76) was not confirmed: both the Southwest and parts of the Southeast have rates below 50 percent in this sample, with the Midlands and North being around 80 and Scotland over 90 percent. Concerning the geographic distribution, the consensus says that negative contraction is more typical for the South, while the North and especially Scotland use auxiliary contraction more often. Tagliamonte & Smith (2002) reject both the general north/south as well as the English/Scottish distinction, noting that no clear cline is visible from their data, and that individual locations both in the North and in Scotland behave contrary to the general pattern. The present analysis integrates both: the Southeast and the Southwest use negative contraction more often, and this decreases toward the North, but there are exceptions to this: the central South and most of the Midlands are closer to the behavior of the North, while Yorkshire, a locus of frequent negative contractions in Tagliamonte & Smith (2002), is closer to the South with regard to this feature. The Midlands are quite interesting here: Shropshire is very different from the dialects toward the East, and has higher probabilities of negative contraction. This is consistent both with the results from Bridge (2006), who found that Derby has high amounts of auxiliary contraction, and with Anderwald's reports of statistically significant differences between areas of the Midlands (2002: 77).

Using the GAM, we can also predict the probability of auxiliary contractions for locations that are not included in FRED. This allows us to compare the results of the model with other studies. Tagliamonte & Smith (2002) is a natural fit. I leave out two of their eight locations – Buckie in the north of Scotland and Culleyback in Northern Ireland – as these

	TIV	HEN	YRK	WHT	MPT	CMN	TYN
actual values	0.43	0.83	0.57	0.98	0.51	1.00	0.87
model predictions	0.50	0.80	0.82	0.82	0.88	0.93	0.87

Table 4.3.: Reported proportions of auxiliary contraction for the locations in Tagliamonte & Smith (2002) and Beal & Corrigan (2005) compared with the model predictions for these locations.

locations lie far outside the range where data is available to the model. Beal & Corrigan (2005) provide comparable percentages from the TLS corpus which contains mostly data from Gateshead in Tyneside. I restrict the model predictions to *to be* as the locus of greatest variability. The probabilities in this model of course depend on the linguistic context on the level of individual tokens, which is not available in enough detail for the other studies. I therefore only predict the most frequent case, i.e. where the preceding word is pronominal (i.e. neither non-pronominal nor *I*) and does not end in a vowel. The effect of vocality is small, and non-pronominal preceding words are relatively rare, both in this corpus and in the examples given in both studies.

Table 4.3 presents the results. For Tiverton (TIV) in the Southwest, Henfield (HEN) in the Southwest, Cumrock (CMN) in Scotland and Gateshead/Tyneside (TYN) in the North, the predicted values are very close to the observed values. For York (YRK), the differences are larger; whereas Tagliamonte & Smith (2002) find auxiliary contraction in only 57 percent of the tokens, the model predicts 82 percent. It should be noted that the locations in Yorkshire that are represented in FRED are rather far away from York; for these dialects, the model predict a slightly lower rate of auxiliary contraction at about 75 percent. Wheatley Hill (WHT) in Durham and Maryport (MPT) in Cumbria exhibit the largest differences. Nevertheless, the match is quite good. For both predictions and local adjustments, the correlation to the observed values lies at about $r = 0.6$. This confirms the correspondence to the findings of previous research, and strengthens the argument that this model is geolinguistically adequate.

To evaluate the implications of this case study on the application of modeling techniques to dialectometric analysis, we first need to quantify the difference between the different values. To do this, I calculate the contribution of the GAM smoother to the final result at each location for both the simple and the full model. The correlation of both models lies at $r = 0.83$; in other words, the results of the simple model explain about 70 percent of the variability in the results of the full model. Considering the numerous differences that exist between the two models, namely different numbers of tokens, different definitions of the variable context, and the fact that only one contains linguistic predictors, this agreement is very high. Furthermore, we can estimate the effect of that last difference by comparing both models to an intermediate one that operates on the restricted data set, but contains no additional predictors. This model has correlation values of around 0.9 to both models, suggesting that the effect of additional predictors constitutes about half of the difference. For the lmer models[6], we find a similar correlation value of 0.77.

The important question, however, is how these values relate to unmodeled values as used in Szmrecsanyi (2013). Concerning proportions, we find that they correlate with both lmer and GAM results on the new data set at $r = 0.43$ and $r = 0.59$. Comparing this to the frequencies of the individual realizations (as they were used in Szmrecsanyi (2013)), we find that auxiliary contraction is slightly negatively correlated ($r = -0.3$), while negative contraction is positively correlated ($r = 0.43$). Together, they explain about as much as the observed proportions. If we assume that either the lmer model or GAM presented in this section is the best available representation of the geolinguistic reality in FRED with regard to this feature, then other measures should be evaluated by how close they are to this representation. And while there is a relation between the unmodeled results and the best models, both simple models come much closer to the best values.

In other words, then, using a much more detailed and linguistically appropriate model and data set leads to results that are different from the simple model, but not overwhelmingly so. This ties in nicely with the argument in Section 3.2.2 that modeling should on average improve the results. We are therefore justified in proceeding with the analysis based on the simple models.

4.1.3. Sociolinguistic summary

The previous section uncovered several reliable effects of the two sociolinguistic predictors, gender and age, as well as their interactions. Here, I will summarize these results and provide some discussion of the patterns. I will only consider predictors that are reliable in both the lmer model and the GAM; marginally significant effects, however, will be included.

[6] For the discussion of the lmer models, the two counties not represented in the new data set had to be removed from consideration, as the model cannot say anything about them. When only GAM values are considered they remain: they are the most likely to exhibit a difference, and removing them would inflate the correlation.

Feature		lmer coef.	GAM coef.
47:	rel. *what*	−1.27	−1.34
40/41:	*don't/doesn't*	−1.15	−1.18
43:	zero aux. progressive	−1.41	−1.08
49:	*as what/than what*	−1.07	−1.04
56/57:	dative alternation	−0.55	−0.69
46:	*wh*-rel.	−0.44	−0.39
6:	*them*	−0.50	−0.38
15:	*to have*	0.15	0.17
13:	*to do*	0.24	0.25
34/35:	contraction with negation	0.30	0.26
24:	*must*	0.27	0.33
25:	*have to*	0.30	0.34
36:	*never*	0.45	0.36
19/20:	habituality	0.39	0.46
51/52:	inf./ger. complementation	0.62	0.64
5:	*us*	0.63	0.66
8/9:	genitive alternation	0.81	0.76

Table 4.4.: Summary of the effects of gender across models, ordered by GAM coefficient. Values below zero indicate lower frequency (or odds for the predicted realization) in female speech.

4.1.3.1. Gender differences

Table 4.4 displays the features where both models found a significant gender difference as well as the model coefficients for that difference. The default level is "male", the most frequent gender in the corpus; the signs of these coefficients therefore indicate the direction in which the female speakers differ. A negative sign indicates lower frequencies, or fewer realizations of the predicted variants for alternations. The table is sorted by the effect of gender in the GAMs. Both models agree on the effect directions and the effect sizes are generally similar, strengthening the confidence in the model results.

These effects can be categorized into several groups. First, there are some core grammatical features that female speakers use significantly more often. These include *us* (Feature 5), the primary verbs *to do* and *to have* (Features 13 and 15) as well as two of the three markers of epistemic or deontic modality, *must* and *have to* (Features 24 and 25). These phenomena may be related to the content of the texts under investigation. Consider the first person plural pronoun *us*. One could speculate that, in oral history narratives, female speakers tend talk more often about groups they are part of, such as their families, which would lead to a higher frequency of first person plural pronouns. This hypothesis can be empirically tested by counting and modeling the other first person plural pronouns *we* and *our*; both GAMs and lmer models result in a significant positive effect for gender. This makes it less likely that the increased frequency of *us* for female speakers is the result of more non-standard usages of *us* as described in Section 4.1.1.1.3. Furthermore, there is evidence that female speakers generally tend to use pronouns more often. In a study tapping the conversational component of the British National Corpus, Rayson et al. (1997) compile a list of words that are particularly characteristic for female speech. Of the 25 words covered there, six are personal pronouns; first person plural pronouns are not included in that list, however. Similarly, Hirschman (1994) found that pronoun usage differed between the male and the female speakers in her (admittedly rather small) sample, with women employing more first and second person plural pronouns than the men did. We will return to difference in pronoun usages in the bottom-up analyses in Section 4.2.3. Moving to the other features in this group, female speakers exhibit a higher usage frequency of all of the surveyed markers of epistemic and deontic modality, and this difference is significant in two out of three cases, *must* and *have to*. The greater frequency of *have to* could partially explain the higher number of tokens of the primary verb *to have*, as one is a subset of the other.

Second, there are many features where female speakers tend to prefer more standard variants. This is not unexpected, as this is hypothesized to be a general pattern of gender differences in language (Chambers 2003). Labov (1990) labels this the Principle I of linguistic change, and it is widely attested around the world with few counterexamples[7]. This tendency toward the standard includes greater likelihood of using the standard *doesn't* with 3rd person singular subjects instead of invariant *don't* (Features 40/41) as well as lower frequencies for the relativizer *what* (Feature 47), for absence of the auxiliary *be* in progressive clauses (Feature 43), for *as what/than what* in comparative clauses (Feature 50) and for *them* followed by potential plural nouns (Feature 6). The only exception to this pattern is *never* as a past tense negator (Feature 36), a non-standard feature that is used more often by women. And this feature can be considered a special case; not only is it widely considered to be a supra-regional feature (cf. Cheshire et al. 1995, Britain 2010), it can be argued that it is historically a rather new development (Lucas & Willis 2012). If this is true and the innovative uses of *never* are still spreading, a higher rate for female speakers would be consistent with Labov (1990)'s Principle II, the fact that in

[7] Auer et al. (2011), for example, show the reverse pattern in their study of southeastern German dialects. They hypothesize that this may be due to female speakers accommodating more to the interviewer's expectations, and thus exhibiting more non-standard or older forms.

most changes, women use the innovative form at a higher frequency.

Then, there are some alternations where both realizations are allowed in Standard English. Compared to men, female speakers are more likely to use the *s*-genitive (Features 8/9), prefer to use *used to* to indicate habituality (Features 19/20), are more likely to contract the negator (Features 34/35), prefer infinitival complementation (Features 51/52), and are more likely to use the prepositional dative after *give* (Feature 56/57). No pattern readily emerges from this. Concerning both habituality and the dative, female speakers tend to prefer the newer variant, while for the genitive and complementation after *begin, start, continue, hate* and *love* they prefer the older form. The gender difference for contraction is called into question by the results of the more complex model in Section 4.1.2, where this difference did not appear. The next section will compare these differences to the effects of age, where appropriate, to see whether there is a apparent-time drift in the present data.

4.1.3.2. Effects of age

Table 4.5 displays the features for which both models found a significant effect of speaker age and their model coefficients. The coefficients are much smaller than in the corresponding table for the effects of gender. This results from the fact that these values are the changes per year of difference between the speaker age and the mean age of all speakers, rather than a single change between two groups. Again, the effect directions in both models are the same and the effect sizes are generally similar. There is one exception to this, the choice of present perfect auxiliary (Feature 22/23). A discussion of this can be found in Section 4.1.1.4.4. It should be kept in mind that an interaction is present in all models and therefore these values are the effects for male speakers. In most cases, female speakers do not behave significantly different; the small number of features where they do can be found in the next section.

As with the gender effects, several feature groups with significant effects can be distinguished. Here interpretation is even more difficult: a given effect can again result from content differences or from actual differences in the grammar, but here these grammatical differences may result from ongoing language change or from performance effects due to the cognitive effects of aging.

First, there is again a group of core grammatical features, albeit much smaller than the corresponding one for gender differences. It again contains a marker of epistemic or deontic modality, *must* (Feature 24), which is less frequent for older speakers, but was used more frequently by women. These results are incompatible with the literature discussed in Section 4.1.1.5 such as Close & Aarts (2010), who note a decrease of *must* in real time.

Feature		lmer coef.	GAM coef.
22/23:	pres. perf. aux.	0.03	−0.05
31:	-*nae*	−0.04	−0.05
56/57:	dative alternation	−0.03	−0.04
46:	*wh*-rel.	−0.03	−0.03
17/18:	future marking	−0.01	−0.02
24:	*must*	−0.01	−0.02
19/20:	habituality	0.01	0.01
36:	*never*	0.01	0.01
30:	nonst. *come*	0.02	0.02
33:	mult. negation	0.02	0.02
44:	nonst. *was*	0.02	0.02
43:	zero aux. progressive	0.03	0.03
51/52:	inf./ger. complementation	0.02	0.03
39:	nonst. verbal -*s*	0.02	0.04
50:	*for to*	0.03	0.04
28:	nonst. weak forms	0.05	0.05
27:	*a*-prefixing	0.07	0.07
32:	*ain't*	0.06	0.07

Table 4.5.: Summary of the effects of speaker age across models, ordered by GAM coefficient. Values below zero indicate lower frequency (or odds for the predicted realization) for each year of speaker age above the average age (75 years).

This is, however, a relatively recent change for British English, and the speakers in FRED may be too old to reflect this shift. Another core feature, *wh*-relativization (46) is used less often as speaker age increases.

Second, there is a large group containing archaic and non-standard forms; as expected, they are more frequently used by older speakers, or, in the case of alternations, have the non-standard variant preferred by older speakers. This group comprises *a*-prefixing (Feature 27), non-standard weak past tense and past participle forms (Feature 28), past tense *come* (Feature 30), *ain't* (Feature 32), multiple negation (Feature 33), *never* as a past tense negator (Feature 36), non-standard verbal *-s* (Feature 39), the absence of the auxiliary *be* in the progressive (Feature 43), non-standard *was* (Feature 44) and unsplit *for to* (Feature 50). There is only one exception to this pattern: the suffix *-nae* (Feature 31), which is used less often by older speakers. As was mentioned in Section 4.1.1.7.1, there are some problems with the values for this feature, as the extremeness of the geographic distribution may have caused modeling problems.

Some of the alternations that exhibited a gender difference again emerge with reliable effects. This group thus consists of choice of habitual marker (Features 19/20), infinitival and gerundial complementation (Feature 51/52) and the dative alternation (Feature 56/57). For all of these, the effect directions for female speakers and for older men are the same: both prefer *used to*, infinitival complementation and the prepositional dative. For older speakers, future marker choice also emerges as significant, with older speakers being less likely to choose *going to* instead of *will* or *shall*. This matches the findings that *going to* as a future marker is still increasing in frequency (Krug 2000, Tagliamonte et al. 2014). There is, however, a complication with the dative alternation, as will be discussed in the next section.

4.1.3.3. Interactions between gender and age

Table 4.6 displays the small number of features where female and male speakers differ in how age affects their linguistic choices. First, there is an interaction such that older women use non-standard verbal *-s* (Feature 39) less often. This contrasts with the effect for older men, who exhibit increased frequencies for this feature, indicating that the frequencies for female speakers remain rather constant through apparent time. For *must* (Feature 24), women show higher frequencies while older speakers use this marker less often; the interaction shows that again there is almost no effect of speaker age for female speakers. *-nae* also reaches statistical significance; see Section 4.1.1.7.1 for the problems with this feature. Finally, the dative alternation emerges as significant. The temporal effect for female speakers goes in the opposite direction than it does for male speakers, leading to

Feature		lmer coef.	GAM coef.
39:	nonst. verbal *-s*	−0.02	−0.05
24:	*must*	0.01	0.02
56/57:	dative alternation	0.06	0.06
3:	*-nae*	0.31	0.25

Table 4.6.: Summary of the interaction of speaker age and gender across models, ordered by GAM coefficient. Values indicate how the effect of age differs for female speakers; lower values indicate lower frequencies or odds for the predicted realization.

a proportional increase of double object datives as speaker age increases. Thus, men and women actually become more similar with age, and the symmetry that was observed in the last section showing that the realization preferred by older speakers is often also the one preferred by female speakers is somewhat broken.

4.1.3.4. Concluding remarks on sociolinguistic results

In summary then, several patterns emerge from the models when looking at the predictors that are (at least marginally) significant in either. First, there is a marked difference between female speakers and older speakers with regard to clearly non-standard features: as expected, women use these at lower rates than men do, while older speakers use them more often. This strengthens confidence in the models. Second, in a number of alternations, women and older speakers prefer the same realization, with older women reversing the trend in the dative alternation.

These results, while suggestive, should not be over-interpreted. First, FRED is not designed to explicitly study sociolinguistic variation (Hernández 2006: 1). Then, the models presented here do not take into account language-internal variables, which are known to influence linguistic choices heavily for the alternations discussed above. For example, Tagliamonte & Lawrence (2000) do not report any effect of age or gender in their study on the determinants of *would/used to* variation, yet find that many other factors, such as animacy or the duration of the habit, influence the choice greatly. Thus, it is entirely possible that differences in frequency or proportions for any feature may not actually be sociolinguistic variation in that feature, but differing input frequencies of such determinants. To make such large-scale, relatively surface-oriented modeling sociolinguistically meaningful, it would need to be supplemented by more careful modeling of at least some of the same data. As in the case study in the previous section, this would require taking other predictors that are known to be relevant into account, to get a sense of whether hidden variables are present that influence the linguistic choices that speakers make, and the degree to which they operate. The analysis of negative and auxiliary contraction has shown that greater care may change the results considerably. Even with that kind of analysis, such a big-picture view will tend to be inaccurate and fuzzy, and will be more useful for hypothesis generation than for proper sociolinguistic analysis. Nevertheless, it is noteworthy that age and gender at least appear to have an effect on many of the features under study here. For the purpose of aggregational geolinguistics, however, it is not crucially important whether the effect of a non-geographic predictor is a sociolinguistically meaningful correlation, or a spurious one resulting from confounding factors. What is important is that neither is directly relevant to dialect geography, and thus accounting for them separately should increase the spatial accuracy of the result.

4.1.4. Geolinguistic summary

4.1.4.1. Feature characteristics

This section presents an overview of the extent to which the distribution of the single features discussed previously is influenced by geography. Unfortunately, there is no simple answer to this question, as various factors play a role. In the single feature discussions, two metrics were provided: first, the variance of the lmer county random effect, indicating how much variability there is in the geographic pattern - the larger this value, the more difference there is between counties. Second, the significance of the GAM smoother - the lower this value, the more the GAM is convinced that the resulting geographic pattern is not just random noise. There are important aspects of the geographic distribution that these values do not cover. One is the complexity of the geographic signal. For example, Feature 24, *must* as a marker of epistemic and deontic modality (Map 14a), has a significant geographic distribution according to the GAM, but the pattern of that distribution is a relatively simple east/west gradient, with higher frequencies in the east. In contrast, Feature 25, *have to* as a marker of epistemic and deontic modality (Map 14b), has a significant distribution as well, but a rather complex pattern: two low-frequency regions in the Scottish Highlands and the central English Southeast, higher frequencies in the English Southwest, the Isle of Man, and Kent, and a somewhat complex pattern of transition regions. One measure that can be used to operationalize complexity of the pattern are the *estimated degrees of freedom* (edf) in the GAM. This measure basically indicates how many different smooth functions the GAM needs, and therefore how different the geographic pattern is from a flat line. High values can result from an overall hilly shape, as in Features 37/38, the *was/weren't* split (Map 20b), or from particularly extreme values in some parts, as in Feature 22, non-standard past-tense *come* (Map 13b).

Another criterion that could be used to evaluate the effect of geography is coherence, i.e. how similar each location is to its neighbors, and how different it is from places further away. An appropriate measure for this is Moran's I (Moran 1950), which consists of a numeric value and a p-value indicating how likely this distribution is to have occurred by chance. These values per feature are provided below for the lmer model, the GAM, and for count-based features for the normalization-based values. As there are 45 tests in total, Bonferroni correction is applied to adjust the customary significance threshold from $\alpha = .05$ to $\alpha = .05/45 = 0.0011$. When the significance value for the feature falls below this number, the value is printed in a bold font in the tables below; the same threshold was applied to GAM smoother significances. Binary weighting was used to determine these values, with a maximum distance of 250 kilometers. Regarding the interpretation of these values, values of I close to zero indicate that the distribution is essentially random, values larger than 0 indicate that closer locations are more similar to each other, and values below 0 indicate that closer locations are more dissimilar to each other.

Table 4.7.: Summary of geographical distribution characteristics. Significant values are highlighted in bold print. Column *var* displays the county random effect variance for the lmer models, larger numbers indicate greater geographic variability. Column *edf* is a measure of GAM complexity, where larger numbers represent a more complex geographic signal. Remaining columns display spatial autocorrelation for lmer models, GAMs, and normalized values. Numbers above zero indicate greater local coherence.

Feature		var	edf	I_{lmer}	I_{GAM}	I_{norm}
1/2:	(non-)st. reflexives	0.742	**18.5**	0.06	0.08	NA
3:	*thee, thou, thy*	2.002	**13**	0.03	0.12	0.02
4:	*ye*	4.029	**28.4**	0.03	-0.06	0.01
5:	*us*	0.213	14.6	0.1	**0.27**	0.03
6:	them	0.427	**18.2**	0.15	**0.44**	**0.2**
7:	synthetic comparison	0.116	2	0.11	**0.69**	0.04
8/9:	genitive alternation	0.192	**17.9**	**0.2**	**0.41**	NA
10:	prep. stranding	0.000	2	0.04	**0.64**	0.04
11/12:	number + *year(s)*	1.386	**28.6**	0.04	0.03	NA
13:	*to do*	0.086	**18.9**	0.11	**0.43**	0.06
14:	*to be*	0.022	**22.1**	**0.39**	**0.49**	**0.37**

Table 4.7.: *(continued)*

Feature		var	edf	I_{lmer}	I_{GAM}	I_{norm}
15:	*to have*	0.034	**18.8**	0.09	**0.26**	0.04
16:	*have got*	0.831	**20.6**	0.13	**0.31**	**0.21**
17/18:	future marking	0.132	**23.6**	0.08	0.07	NA
19/20:	habituality	1.113	**27.4**	0.12	**0.34**	NA
21:	progressive	0.305	**14.2**	0	0.12	0.11
22/23:	pres. perf. aux.	3.788	**27.2**	-0.02	-0.06	NA
24:	*must*	0.144	2.2	0.07	**0.39**	0.05
25:	*have to*	0.172	20.1	0.01	0.08	-0.04
26:	*got to*	0.694	**18**	**0.4**	**0.57**	**0.38**
27:	*a*-prefixing	3.568	**26.5**	0.01	-0.01	0.07
28:	nonst. weak forms	0.915	28.1	-0.04	-0.06	-0.04
29:	nonst. *done*	1.764	**13.4**	**0.33**	**0.65**	**0.38**
30:	nonst. *come*	0.591	**28**	0.09	**0.25**	**0.28**
31:	*-nae*	90.463	16.1	**0.29**	**0.38**	**0.29**
32:	*ain't*	2.538	**13.1**	**0.43**	**0.6**	**0.51**
33:	mult. negation	1.028	**18.3**	**0.4**	**0.61**	**0.4**
34/35:	contraction/negation	0.505	**20.8**	0.16	**0.34**	NA
36:	*never*	0.130	7.5	-0.05	**0.16**	-0.06
37/38:	*wasn't/weren't*	0.590	**27.2**	-0.11	-0.08	NA
39:	nonst. verbal *-s*	1.031	**20.3**	-0.04	-0.12	-0.04
40/41:	*don't/doesn't*	5.261	**11.7**	**0.36**	**0.74**	NA
42:	*there is*	0.151	**14.7**	**0.21**	**0.45**	0.06
43:	zero aux. progressive	0.473	8.9	-0.08	**0.25**	-0.04
44:	nonst. *was*	0.303	12.3	-0.05	0.04	-0.08
45:	nonst. *were*	3.220	**18.1**	-0.03	0.12	-0.05
46:	*wh*-rel.	0.412	11.6	-0.01	**0.22**	-0.02
47:	rel. *what*	0.890	**9.8**	0.14	**0.53**	0.01
48:	rel. *that*	0.153	**11.2**	0.02	**0.33**	0.01
49:	*as what*	0.181	9.8	-0.1	-0.03	0.11
50:	*for to*	1.703	**18.6**	0	0.1	-0.07
51/52:	inf./ger. compl.	0.924	**21.8**	-0.11	-0.05	NA
53/54:	zero/*that* compl.	0.236	**20.6**	0.08	0.16	NA

Table 4.7.: *(continued)*

Feature		var	edf	I_{lmer}	I_{GAM}	I_{norm}
55:	lack of inversion	0.679	10.8	-0.06	-0.05	0
56/57:	dative alternation	0.061	8.5	-0.01	**0.35**	NA

Table 4.7 provides an overview of the above measures for all features.

I will now give a brief summary of the results for each measure. Table 4.8 displays the top 10 features according to the lmer county random effect. The list contains several features that are well-known for their strong geographic distribution, for example Features 40/41, *don't* as opposed to *doesn't* with 3rd person singular subjects, Feature 32, the negator *ain't*, and Feature 29, non-standard past tense *done*, all of which are much more frequent in Southern British English than in Northern and Scottish dialects. The list also contains Features that are overall very rare and appear only in certain regions, such as Feature 31, the negating suffix *-nae*, which is mostly restricted to Scotland, as well as Feature 27, *a*-prefixing on *-ing* forms, and Features 22/23, the present perfect auxiliary *be* as opposed to *have*, which are largely restricted to Suffolk in East Anglia. Of the remaining features, Feature 45, non-standard *were*, is particularly frequent in parts of the Midlands and the North of England and Feature 3, the archaic pronouns *thee, thou, thy*, shows a partial east/west split. Features 4, the archaic pronouns *ye*, and 50, unsplit *for to*, complete the list; both appear primarily in Scotland, with Feature 50 also having higher frequencies in the English south.

Table 4.9 shows the features with the highest and lowest GAM smoother estimated degrees of freedom, restricting our attention to those where the geographic distribution significantly adds to the model quality. More than 50 percent of the top six are alternations, suggesting that alternations tend to have rather complex distributions. Most of these involve hilly patterns where areas of high and low frequency may be relatively close to one another, and that therefore eschew easy summarization. The same is true for the non-alternations on the list, the archaic pronoun *ye* (Feature 4) and non-standard past tense *come*. Regarding the bottom features, we find one of the features from table 4.9, *don't* or *doesn't* with 3rd person singular subjects (Feature 40/41). While the geographic differences here are quite large, they follow a relatively simple north/south pattern. All features related to relativization are similarly simple and are mostly centered around one region, decreasing as one moves away from there. The by far simplest feature, the marking of epistemic and deontic modality using *must* (Feature 24), is hardly more than

Feature		var_{lmer}
31:	-*nae*	90.463
40/41:	*don't/doesn't*	5.261
4:	*ye*	4.029
22/23:	pres. perf. aux.	3.788
27:	*a*-prefixing	3.568
45:	nonst. *were*	3.220
32:	*ain't*	2.538
3:	*thee, thou, thy*	2.002
29:	nonst. *done*	1.764
50:	*for to*	1.703

Table 4.8.: Largest lmer county effect variances. Larger numbers indicate greater geographic variability.

Feature		edf_{GAM}
11/12:	number + *year(s)*	**28.6**
4:	*ye*	**28.4**
30:	nonst. *come*	**28**
19/20:	habituality	**27.4**
22/23:	pres. perf. aux.	**27.2**
37/38:	*wasn't/weren't*	**27.2**
44:	nonst. *was*	12.3
40/41:	*don't/doesn't*	**11.7**
46:	*wh*-rel.	11.6
48:	rel. *that*	**11.2**
47:	rel. *what*	**9.8**
24:	*must*	2.2

Table 4.9.: Largest and smallest significant GAM effective degrees of freedom. Larger numbers represent a more complex geographic signal.

a linear gradients from the east to the west. While this distribution is significant (without Bonferroni correction), it should be kept in mind that the corresponding model does not fit the data particularly well.

Let us now move from model characteristics the model predictions, and test for which features geographically close areas are also linguistically close, using spatial autocorrelation. There are quite a few Features where Moran's I is significant for the lmer models;

Feature		I_{lmer}
32:	*ain't*	**0.43**
26:	*got to*	**0.4**
33:	mult. negation	**0.4**
14:	*to be*	**0.39**
40/41:	*don't/doesn't*	**0.36**
29:	nonst. *done*	**0.33**
31:	-*nae*	**0.29**
42:	*there is*	**0.21**
8/9:	genitive alternation	**0.2**
34/35:	contraction with negation	0.16

Table 4.10.: Highest significant Moran's *I* values for lmer county predictions. Numbers above zero indicate greater local coherence.

Table 4.10 displays those with the strongest geographical association. First, however, consider the full list in Table 4.7 and compare them to those for the normalization-based values, where possible. There are three cases where the normalized counts reach the threshold, but the lmer predictions do not. The include *them* after potential plural nouns (Feature 6), *have got* as a marker of possession (Feature 16), and non-standard uses of *come* (Feature 30). In contrast, one feature is now significant: plural *there is/was* (Feature 42). The lmer predictions therefore do not reflect geographic signal quite as strongly. Partially, this is related to the admittedly strict Bonferroni correction. When not correcting for multiple comparisons the lmer values fare much better: now the lmer values reach significance in six cases where the normalized values do not, and only miss three where they do. Let us now turn to the table at hand; there are nine cases where a geographic pattern is apparent. This list contains the already familiar Southern British features, as well as three new ones with a similar distribution: multiple negation (Feature 33), *got to* (Feature 26) and to a lesser degree the genitive alternation (Features 8/9). Then, there is a group of features that are more prevalent in Scotland and the North of England: the already familiar features -*nae* and *to be*, as well as *there is/was* with plural subjects.

There are many more features where Moran's *I* is significant for the predicted smoother GAM values. In fact, the GAM confirms all features that are significant for the unmodeled values, and adds nine new ones. Table 4.11 shows the top ten features with the highest spatial autocorrelation. The list contains mostly features that are familiar from the earlier lists, with three new entries: First, the relative particle *what* (Feature 47) joins the features

Feature		I_{GAM}
40/41:	*don't/doesn't*	**0.74**
7:	synthetic comparison	**0.69**
29:	nonst. *done*	**0.65**
10:	prep. stranding	**0.64**
33:	mult. negation	**0.61**
32:	*ain't*	**0.6**
26:	*got to*	**0.57**
47:	rel. *what*	**0.53**
14:	*to be*	**0.49**
42:	*there is*	**0.45**

Table 4.11.: Highest significant Moran's I values for GAM predictions. Numbers above zero indicate greater local coherence.

that are most common in the south. Then there are two features where the GAM smoother did not reach significance, synthetic adjective comparison (Feature 7) and proposition stranding (Feature 10). Their presence on this list is related to that: as the GAM only finds little evidence of a geographic distribution, and that distribution is very linear, close locations are by necessity similar to one another. As these features are relatively rare, they do not influence the overall result by much.

In summary then, compared to the normalization-based values, the lmer predictions are slightly more conservative, although less so if the restrictive significance thresholds are relaxed, and the GAM predictions seem anti-conservative for features with few attestations and no clear signals.

4.1.4.2. The geographic relations between features

So far, this section has discussed features individually by the characteristics of their geographic distribution. In this section, I tackle the question of how similar the distributions between the features are. The process here is related to the aggregational analyses that will be presented in Chapter 5, but with some crucial differences. First, instead of classifying dialects by the features they exhibit, I classify features by their distribution across dialects. This necessitates two major modifications to the process. The starting point is the same: the predictions made by the lmer and GAM models after the application of the logarithmic transformation and after enforcing minimum frequencies or odds. When using the binary alternations to classify dialects, only the odds for the predicted realization are included. Here, both the values for the predicted realization as well as those of

Cluster	Heb	MAN	Mid	N	ScH	ScL	SE	SW	Wal
light blue	−0.99	−0.56	0.27	−0.13	−1.03	0.19	0.46	0.06	−0.37
dark blue	−0.94	0.39	0.76	0.14	−0.95	−0.70	0.45	0.71	0.40
red	−0.12	0.02	−0.60	0.14	0.31	0.74	−0.51	−0.74	−0.32
dark red	1.62	0.19	−0.35	−0.20	0.80	−0.22	−0.14	0.18	0.27

Table 4.12.: Associations of feature bundles to FRED regions. Values close to zero indicate that features from that bundle are overall distributed in that region as in the whole corpus; higher and lower values indicate more and less frequently used bundles.

the alternative realization are included. The reason for this is as follows: consider, for example, the alternation between *don't* and *doesn't* with third person singular subjects. *Don't*, the predicted realization, is a feature of the English south, and we would therefore expect it to group with other features that are more frequent there. The choice of the realization to predict is, however, arbitrary, and we could have chosen *doesn't* just as well. In that case, we would expect the feature to group with things that are rare in the South. By including both variants, each pattern will be apparent. The alternate realizations are marked by the addition of "non-default" to their label. The second crucial difference to the classification of dialects is the scaling of values. Some features are very frequent, such as the primary verbs, while others are quite rare. The goal here is to identify those that have a similar spatial distribution, but the overall feature frequency obscures this: features that are relatively frequently used everywhere will be similar to each other, even if their distribution patterns match other features more closely. Therefore, the values are first scaled (i.e. divided by their standard deviation) and centered around 0. This puts all features on the same scale. After these adjustments, the aggregation proceeds in the usual fashion, using the Euclidean distance measure.

Figure 4.1 shows the result of hierarchical clustering on the resulting data set. Four groups are highlighted in the dendrogram. The major split between groups is the one separating the light and dark blue colors from the red ones. We can investigate the associations between these feature clusters and the regional classifications in FRED by averaging the value of the features in each cluster per region. A value of 0 would then indicate that, with regard to this cluster of features, the varieties exhibit about average frequencies. A value of 1 would indicate that these features are overall one standard deviation more frequent in that area than they are in the whole corpus.

Table 4.12 displays the result. Let us begin with the blue colors. The dark blue cluster contains the most distinctive features of the English south, especially the Southwest, very closely together, from non-standard *done* to multiple negation. Also included in this cluster are several features that a primarily characterized by being rare in Scotland, such as *to have*, *us* or the dative alternation. The light blue cluster, on the other hand, contains features that appear mostly in the Southeast, such as the relativizer *what*, the double object dative, or non-standard *was*. The light red group contains features that are either distinctively rare in (especially southern) English English, such as alternate realizations of features in the blue clusters, and in particular features distinctive for the Scottish Lowlands, such as -*nae*, *to be*, or *there is* with plural subjects. The final cluster contains features that are particularly frequent in the Scottish Highlands and the Hebrides, such as the *going to* future, the progressive, or explicit plural marking on *years* after numerals. Note that this cluster, particularly prevalent in comparably young dialects, contains many standard realizations in alternations, such as standard reflexives and the present perfect using the auxiliary *be*. This is consistent with the results from Szmrecsanyi (2013: 84ff.), who finds that these varieties are most similar to both British and American Standard English.

Several comments can be made regarding the feature groups that emerge. First, features that have elements in common tend to group together. For example, the features involving forms of the primary verb *to do*, such as non-standard *done* or invariant *don't*, group very closely together. The same is true for *to have* and *have to* as well as for *have got* and *got to* (all in the dark blue cluster); this dovetails nicely with Schulz (2012)'s hypothesis that

> an intraferential process involving HAVE$_{\text{poss}}$, HAVE GOT and HAVE TO can be postulated, where the co-presence of the possessive expressions HAVE$_{\text{poss}}$ and HAVE GOT on the one hand and a sharp rise in the frequency of HAVE TO on the other hand motivate an extension of the subcategorization frame of HAVE GOT from possessee NP to to-infinitival complements.

Of course there are concerns about circularity between some of these features; for example, all instances of *have got* and *have to* also count toward the overall frequency of *to have*. Similar considerations apply to the forms of *to do*, and (in the light blue cluster) to non-standard *was*, which is closely linked to non-standard verbal -*s* and *wasn't*, the alternate realization of *wasn't/weren't*. Such objections can be ruled out, however, in another case: the distribution closest to those of relativizer *that* is explicit complementation using *that*. The frequencies for these two features are not included based on the same tokens, and these tokens were manually checked to rule out wrong classifications. This suggests structural persistence of the type that Szmrecsanyi (2005; 2006) calls β-persistence, the idea that the linguistic choices of speakers are “affected by non- variable linguistic patterns that share structural, lexical, or other characteristics with one of the choice options” (2005: 140). A further, albeit less clear example for this involves gerundial complementation. Szmrecsanyi (2006) notes that recency of the last -*ing* form in general (i.e. including progressive forms) increases the probability of a speaker to use gerundial complementation. In the dendrogram, this is the alternate realization of Features 51/52, and, while they are not as close to one another as the previous examples, it appears in the same cluster as the two features counting progressives (Features 21 and 43).

Cluster analyses of this data in the usual direction, i.e. grouping the dialects according to their features, will proceed in Chapter 5. The remainder of this chapter will discuss the results of the bottom-up syntactic analysis.

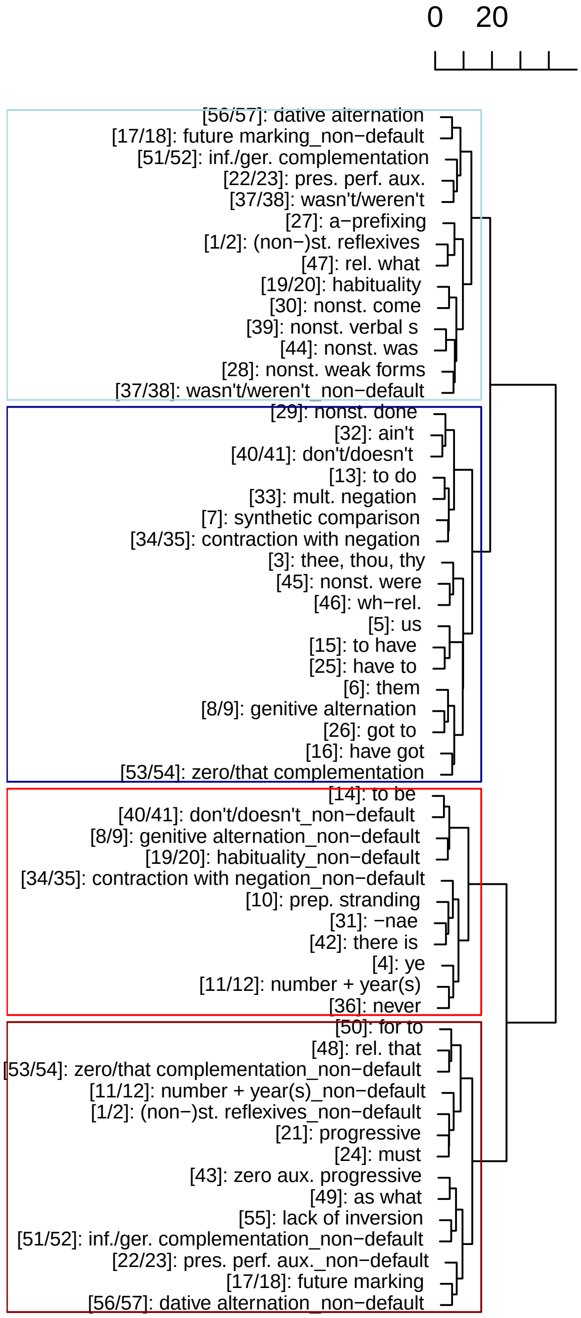


Figure 4.1.: Aggregate view on the feature distribution: Hierarchical cluster plot

4.2. POS n-gram-based analyses

This section discusses the results of the method introduced in Section 3.2.3 on the part-of-speech-tagged FRED-S corpus. Only results pertaining to individual n-grams and their distribution will be covered here, the results of aggregation and hierarchical clustering will be provided in Sections 5.2.4 and 5.2.5. This section will proceed as follows: after a basic overview, a selection of the unigrams and bigrams that have emerged as particularly relevant in their geographic distribution will be given. Here, geographic distribution refers to differences across counties; whether geographically close counties are also linguistically similar will not be considered. Then, a similar approach will be used to find which n-grams vary along the sociolinguistic axes, gender and age.

In total, the corpus consists of 1,008,213 unigrams of 225 types, and 943,541 bigram tokens, spread over 9,035 different types.

Comparing the two distinctiveness measures, p-distinctiveness and r-distinctiveness lead to very similar results, with Spearman's rank correlation coefficient being $\rho = -0.8$. Of the two, p-distinctiveness prefers frequent bigrams, with the rank correlation to the bigram frequency being $\rho = 0.8$ (r-distinctiveness: $\rho = 0.55$).

The complete list of POS tags can be found in Appendix A.

4.2.1. Geolinguistic results: unigrams

Table 4.13 displays the unigrams that have the highest total r-distinctiveness. Several of these harbor dialectologically relevant phenomena, and will be discussed in greater detail with the bigrams below. The first is *was/were* variation, represented in the table by *were* (`VBDR`), and, to a certain degree, *there* (`EX`), discussed in existential/presentational contexts in Section 4.2.2.1. `VMK`, *used* in the habitual marker *used to*, is somewhat lower on the r-distinctiveness scale (rank 25), but is the second-highest on the p-distinctiveness scale; its discussion can be found in Section 4.2.2.2. `PPHO2`, *them*, is related to the use of *them* as a plural determiner, illustrated in Section 4.2.2.3. Plural nouns (`NN2`) hide, through misclassification, the Scottish negator *-nae*, as will be shown in Section 4.2.2.4. Several forms of *do*, including the unmarked form (`VD0`) that is included in the list of top unigrams, are involved in dialectologically relevant phenomena, which will be the topic of Section 4.2.2.5.

ngram	example	rel.	rank.rel.	n.sig.	rank.n.sig	N
PPHS1	*he*	0.54	1	95	14	18638
VBDR	*were*	0.66	2	107	4	6317
NNB	*Mr*	0.78	3	91	20	1568
RL	*here*	0.88	4	106	5	11408
RP	*over*	0.94	5	102	9	21894
EX	*there*	0.94	6	89	22	5855
RT	*then*	1.06	7	85	30	11964
VBZ	*'s*	1.06	8	98	12	9511
NNL1	*Street*	1.14	9	89	23	1228
PPHO2	*them*	1.38	10	88	25	6740
NP1	*Tom*	1.48	11	102	8	17706
VVI	*instruct*	1.52	12	80	37	31465
II	*at*	1.64	13	85	29	47539
CC	*and*	1.66	14	71	49	43320
CST	*that*	1.70	15	73	46	3819
CSA	*as*	1.76	16	85	28	2020
RR21	*sort*	1.88	17	93	17	3116
RR22	*course*	1.90	18	93	18	3107
VD0	*do*	1.90	19	81	35	3480
PPY	*you*	1.94	20	93	16	24368

Table 4.13.: Most relevant unigrams. Column *ngram* contains the POS tag, column *example* an example. Column *rel* displays the reliability score (lower is more noteworthy), column *n.sig* the number of pairwise significant differences (higher is more noteworthy). Rank columns indicate the rank of this unigram in the total list when ordered by that metric. Final column shows total unigram frequency.

4.2.2. Geolinguistic results: bigrams

This section provides examples for statistically and geolinguistically relevant bigrams and discusses their distribution.

4.2.2.1. *was*/*were* variation

Variation between *was* and *were* played an important part in the model-based approach to CBDM, most importantly in Features 37/38 (Section 4.1.1.7.6), 44 and 45 (Sections 4.1.1.8.5f.). As such, it is reassuring that several bigrams with particularly high relevance involve either *was* (tag: `VDBZ`) or *were* (tag: `VBDR`). Table 4.14 shows the 20 most relevant bigrams involving either tag; and (41) to (43) show example realizations of the top bigram patterns. Many or these involve a combination of tags that are either ungrammatical in Standard English (41), or the corresponding standard form (42):

(41) Non-standard *was*/*were*:

a. `PPH1.VDBR`:
And that's how it $_{\texttt{PPH1}}$ were_`VBDR` kept going . [LAN_003]

b. `PPHS1.VDBR`:
He_`PPHS1` were_`VBDR` a good mam and dad , yeah . [YKS_004]

c. `PPHS2.VBDZ`:
They , they_`PPHS2` was_`VBDZ` both born in Preston . [LAN_005]

d. `PPIS1.VBDR`:
I_`PPIS1` were_`VBDR` born in 1917 . [NTT_004]

e. `NN1.VBDR`[8]:
So the eh Manager_`NN1` were_`VBDR` going through and he said [. . .] [WIL_001]

(42) Standard *was*/*were*

a. `PPHS1.VDBZ`:
He_`PPHS1` was_`VBDZ` a tackler ... [YKS_011]

b. `PPHS2.VBDR`:
Aye , they_`PPHS2` were_`VBDR` Tyne Corps . [NBL_007]

Are these bigrams in competition? To test this, we can compare the attested patterns with *was* to those containing *were* using Spearman's rank correlation coefficient. The overall mean correlation is 0.03, indicating that in general there is no competition between was and were. Individual bigrams, however, show a strong negative correlation, crucially including most of the patterns in Table 4.14 and the examples in (41) and (42). The only

[8] Note that this tag combination is not always non-standard, e.g. *What part_ **NN1** were_ **VBDR** you playing?* [Interviewer in ELN_011]; I count this bigram as non-standard as most attestations by informants are clearly non-standard.

ngram	example	rel.	rank.rel.	n.sig.	rank.n.sig	N
PPH1.VBDR	*It were*	0.10	1	66	127	431
PPHS1.VBDR	*she were*	0.28	2	65	133	330
PPHS2.VBDZ	*they was*	0.48	3	92	12	779
PPHS2.VBDR	*they were*	0.68	6	90	15	1724
EX.VBDZ	*there was*	1.24	16	89	17	3110
PPIS1.VBDR	*I were*	1.26	17	61	180	300
EX.VBDR	*there were*	1.34	19	87	22	633
VBDR.AT1	*were a*	1.70	24	62	176	350
PPIS2.VBDZ	*We was*	1.77	26	65	136	353
NN1.VBDR	*War were*	1.78	28	58	215	231
PPIS2.VBDR	*we were*	1.80	30	99	5	835
VBDR.RG	*were about*	1.96	38	67	123	271
PPY.VBDZ	*you was*	2.22	57	56	240	298
VBDR.JJ	*were little*	2.26	58	62	177	660
PPIS1.VBDZ	*I was*	2.30	62	62	169	2662
VBDR.RR	*were always*	2.34	65	67	124	471
PPHS1.VBDZ	*he was*	2.36	69	87	24	2822
VBDR.DB	*were all*	2.49	82	56	241	210
PPY.VBDR	*you were*	2.58	88	86	26	462
PPH1.VBDZ	*it was*	2.62	94	93	9	4614

Table 4.14.: Most relevant *was/were* related bigrams. Column *ngram* contains the POS bigram, column *example* an example. Column *rel* displays the reliability score (lower is more noteworthy), column *n.sig* the number of pairwise significant differences (higher is more noteworthy). Rank columns indicate the rank of this bigram in the total list when ordered by that metric. Final column shows total bigram frequency.

exception to this is PPIS1.VBDZ/VBDR, *I was/were*, which shows no correlation ($\rho = 0.02$). Furthermore, the bigrams listed there tend to be among those with the overall strongest negative correlation: PPIS2.VBDZ/VBDR (*we was/were*, $\rho = -0.71$), PPHS1.VBDZ/VBDR (*she was/were*, $\rho = 0.58$) and PPHS2.VBDZ/VBDR (*they was/were*, $\rho = -0.56$) top the list. Maps 30a and 30b illustrate this competition for PPIS2.VBDZ/VBDR variation. The non-standard form *we was* 30b is particularly frequent in the Southeast of England, especially London and Kent. The standard form *we were* is particularly frequent in the North of England and especially the Lothians. The Southwest, finally, shows intermediate frequencies for both. The former clearly shows a similarity to the result of the modeled feature counts in Map 23b (page 125).

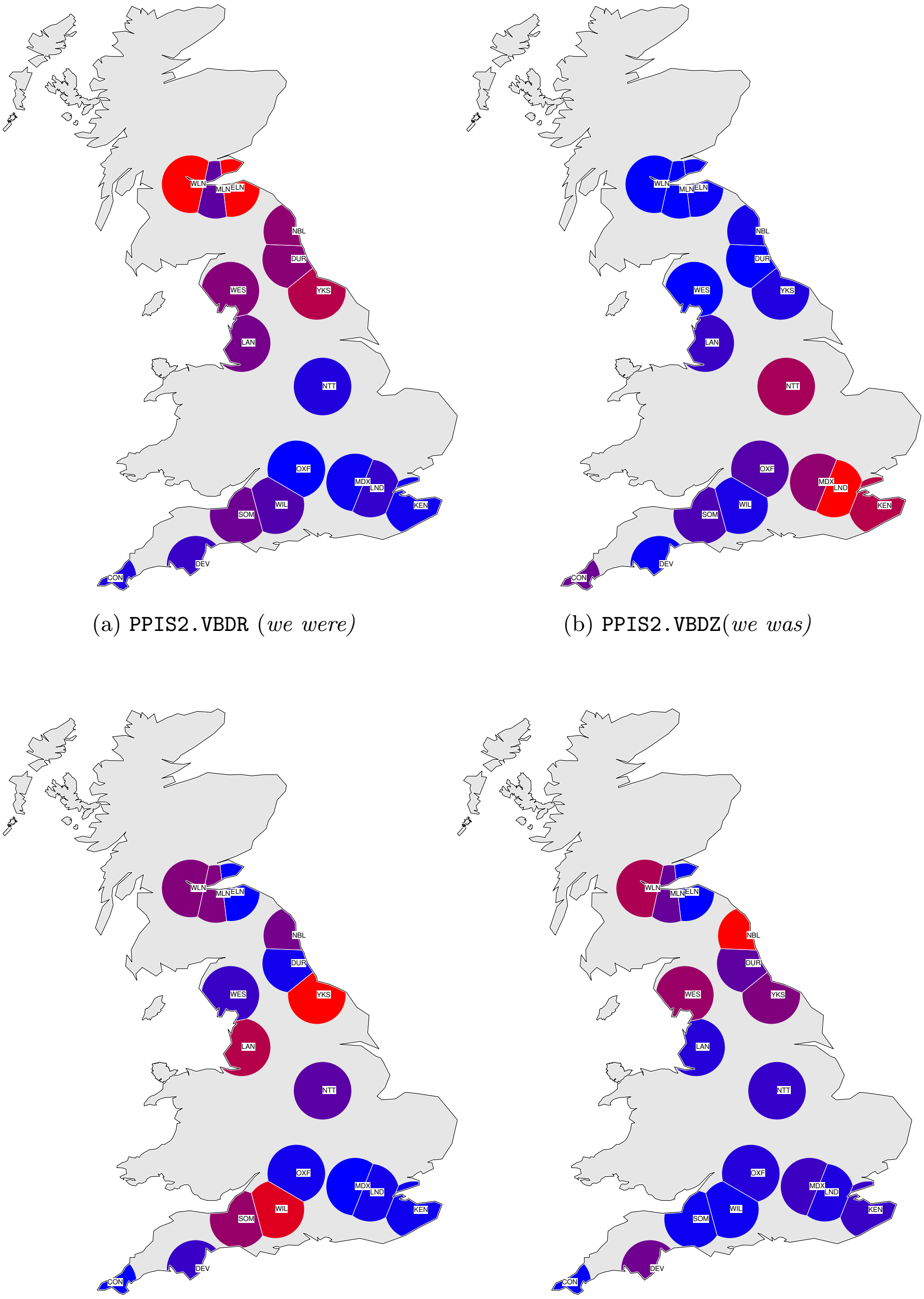
(43) Existential *there was*/*there were*

a. There_EX were_VBDR three different parts. [LAN_003]
b. There_EX were_VBDR Roseley Camp and there_EX were_VBDR Brockton Camp . [YKS_010]
c. And there_EX was_VBDZ a pump . [YKS_009]
d. And there_EX was_VBDZ four girls [...] [LND_003]

A special case is existential *there* followed by either *was* or *were*. As the examples in (43) show, both patterns can involve standard or non-standard uses. The special case of existential *there was* with a plural subject (as in (43d)) is part of Feature 42 in Szmrecsanyi's list, and is associated with Scotland (see Map 22a on page 120). We might expect that a high prevalence of this feature would lead to more instances of *there was* and fewer of *there were*. Calculating the correlation between usage frequencies shows that not only is there no competition, both patterns are modestly positively correlated ($rho = 0.28$). In other words, in counties where speakers use existential sentences involving *was* more often, they also use existential sentences with *were* more often, and alternation between *was* and *were* does not seem to have a strong effect on the distribution. Maps 30c and 30d illustrate this: many counties end up with similar colors; with *there was* and *there were* both being more frequent in the North of England and Scotland. This suggests that Feature 42 is confounded by the frequency of existential or presentational *there*-constructions, which may account for the divergence from the WAVE results observed in Section 4.1.1.8.3.

4.2.2.2. *used to*

Another POS tag that appears in patterns high on the distinctiveness scales is VMK (see Table 4.15), indicating one of the modal catenatives *used* or *ought*. In practice, the vast



(a) `PPIS2.VBDR` (*we were*)

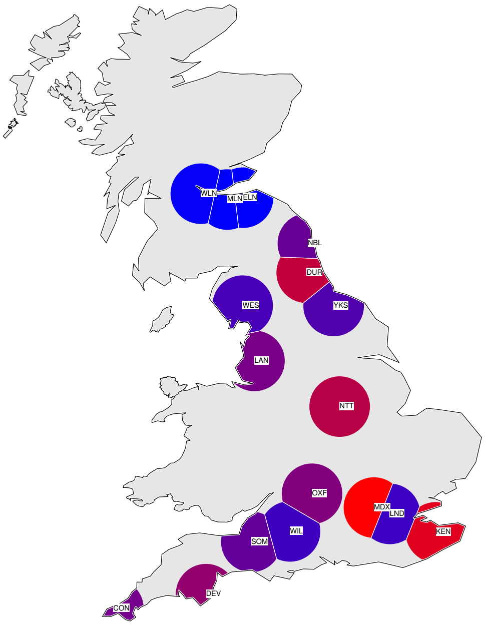
(b) `PPIS2.VBDZ`(*we was*)

(c) `EX.VBDR` (*there were*)

(d) `EX.VBDZ` (*there was*)

Map 30: Visualization of geographic variation involving *was*/*were*. More reddish colors indicate greater frequency of the bigram in question.

majority of instances involve specifically *used to* (see examples in (44)).



Map 31: Visualization of geographic variation involving *used to* (`VMK.TO`). More reddish colors indicate greater frequency of this bigram.

(44) a. `NN1.VMK`
My grandmother_`NN1` used_`VMK` to wear one of those . [DEV_001]

b. `RR.VMK`
Oh it always_`RR` used_`VMK` to be Teignmouth . [DEV_005]

c. `PPIS2.VMK`
We_`PPIS2` used_`VMK` to go to church . [NTT_004]

The vast majority of relevant bigrams are different subject types, and a correlation analysis shows that overall, all bigram frequencies involving `VMK` correlate rather strongly with the frequency of `VMK.TO` (mean $\rho = 0.25$); this holds especially for the most distinctive bigrams such as `NN1.VMK` ($\rho = 0.90$), `RR.VMK` ($\rho = 0.89$) or `PPIS2.VMK` ($\rho = 0.69$). This strongly suggests that these bigrams largely measure the same thing, namely the frequency of *used to* as a habitual marker. Map 31 displays the geographic distribution of `VMK.TO`. This pattern is particularly frequent in Middlesex and Kent in the Southeast and particularly rare in Scotland. The Southwest and the North of England as well as London show various degrees of intermediate frequencies.

Used to as a habitual marker is included in the model-based analysis in alternation with the habitual marker *would* as Feature 19/20, and the corresponding Map 12b can

ngram	example	rel.	rank.rel.	n.sig.	rank.n.sig	N
NN1.VMK	*man used*	0.68	5	88	20	760
RR.VMK	*always used*	1.04	13	69	109	368
VMK.TO	*used to*	1.74	25	108	1	11223
PPIS2.VMK	*We used*	1.78	27	78	58	2187
PPHS2.VMK	*They ought*	1.94	35	88	21	2092
PPHS1.VMK	*he used*	2.06	43	95	7	1539
EX.VMK	*there used*	2.48	79	63	155	396
PPY.VMK	*you used*	2.78	112	58	219	481
CST.VMK	*that used*	3.12	152	53	267	203
NP1.VMK	*White used*	3.55	212	49	310	156

Table 4.15.: Most relevant related bigrams related to *used to*. Column *ngram* contains the POS bigram, column *example* an example. Column *rel* displays the reliability score (lower is more noteworthy), column *n.sig* the number of pairwise significant differences (higher is more noteworthy). Rank columns indicate the rank of this bigram in the total list when ordered by that metric. Final column shows total bigram frequency.

be found on page 94. Here, the match between models and the n-gram methods is a bit worse; while both place a peak of *used to* frequency in the Southeast, the models do not agree with the feature's rarity in Scotland.

4.2.2.3. *them*

Table 4.16 shows the distinctive bigrams related to *them* (tag: PPHO2). Of these only one ranks particularly high: PPHO.NNT2, i.e. *them* followed by a temporal plural noun. Almost all of these attestations involve the particular non-standard idiom *them days* (45a), with a small number of other temporal nouns also attested, namely *years* (45b), *hours* and *times.*

(45) bigrams involving *them*

a. PPHO.NNT2
   And there was no combines in them_PPHO2 days_NNT2 . [YKS_009]

b. PPHO.NNT2
   With all them_PPHO2 years_NNT2 being out one ? [NTT_013]

c. II.PPHO2
   And the great bank around_II them_PPHO2. [DEV_001]

d. PPHO2.NN2
   Them_PPHO2 farmers_NN2 was rationed for meat , [. . .] [SOM_014]

ngram	example	rel.	rank.rel.	n.sig.	rank.n.sig	N
PPHO2.NNT2	*them days*	0.94	9	93	10	527
II.PPHO2	*in them*	2.76	107	79	50	1079
VVI.PPHO2	*pass 'em*	3.10	151	61	184	1516
VVD.PPHO2	*turned them*	3.44	192	68	115	694
VVN.PPHO2	*put them*	3.48	199	25	854	147
DB.PPHO2	*all them*	3.57	217	28	729	80
PPHO2.II	*them from*	3.78	262	52	283	609
PPHO2.CC	*them and*	3.94	306	23	908	401
PPHO2.JJ	*them straight*	4.05	327	28	743	148
PPHO2.PPHS1	*them he*	4.37	414	10	1663	39
PPHO2.VVI	*them keep*	4.39	420	20	1035	100
PPHO2.RR	*them properly*	4.91	600	27	767	174
PPHO2.CCB	*them but*	5.19	716	8	1855	69
VVG.PPHO2	*taking them*	5.38	800	18	1156	202
PPHO2.RP	*them up*	5.40	804	47	340	770
VV0.PPHO2	*call them*	5.50	862	36	535	988
IO.PPHO2	*of them*	5.62	918	38	486	843
VHI.PPHO2	*have them*	5.71	980	14	1373	68
PPHO2.VVD	*them came*	5.72	983	5	2289	75
PPHO2.NN2	*them berths*	5.84	1042	36	532	268

Table 4.16.: Most relevant bigrams related to *them*. Column *ngram* contains the POS bigram, column *example* an example. Column *rel* displays the reliability score (lower is more noteworthy), column *n.sig* the number of pairwise significant differences (higher is more noteworthy). Rank columns indicate the rank of this bigram in the total list when ordered by that metric. Final column shows total bigram frequency.

e. PPHO2.NN2
We call 'em_PPHO2 Cats_NN2 , see . [SOM_002]

Other bigrams involving *them* are less clear in what precise dialectal feature they reflect. For example, it is not immediately clear why the distribution of a preposition (tag: II) followed by *them* as in (45c) has comparatively strong geographic differences, except that many of the PPHO2.NNT2 cases are preceded by a preposition. One pattern with a rather straightforward dialectological interpretation, however, is *them* followed by a general plural noun (tag: NN2). Most instances of this pattern involve *them* in the role of demonstrative *those*, as in (45d). It should be noted, however, that the same pattern can also appear in standard syntactic contexts, such as in (45e). The distribution of PPHO2.NN2 is very similar to that of PPHO2.NNT2 as measured by Spearman's rank correlation coefficient ($\rho = 0.71$). This indicates that the use of *them* as a demonstrative is, at least in this data set, especially distinctive in a particular, limited context: temporal nouns such as *days* or *years*.

Maps 32a and 32b display the geographical distribution of *them* followed by temporal and general plural nouns, respectively. Both patterns are especially rare in Scotland, and rather frequent in the North of England, Kent, and Oxfordshire.

Them, restricted to nouns likely to be plural, is included as Feature 6 in the model-based analyses. Map 7 on page 85 shows the result, which is essentially the same as for the bigrams above: higher frequencies in the North and in central England, lower frequencies in the Southwest and in Scotland,

4.2.2.4. *-nae*

Some bigrams involving a pronoun, especially the first person singular subject pronoun *I* (tag: PPIS1), followed by a plural noun (tag: NN2) emerge as both significant and distinctive. Closer inspection shows that these often involve incorrectly tagged words. More specifically, verbs with the suffix -*nae*, which is used in Scottish dialects to negate modal auxiliaries and *do* (cf. Section 4.1.1.7.1), are incorrectly tagged as plural nouns by CLAWS. Presumably, this is because English words ending in nae are usually plural forms of nouns of Latin origin, for example *antennae* being the plural of *antenna*.

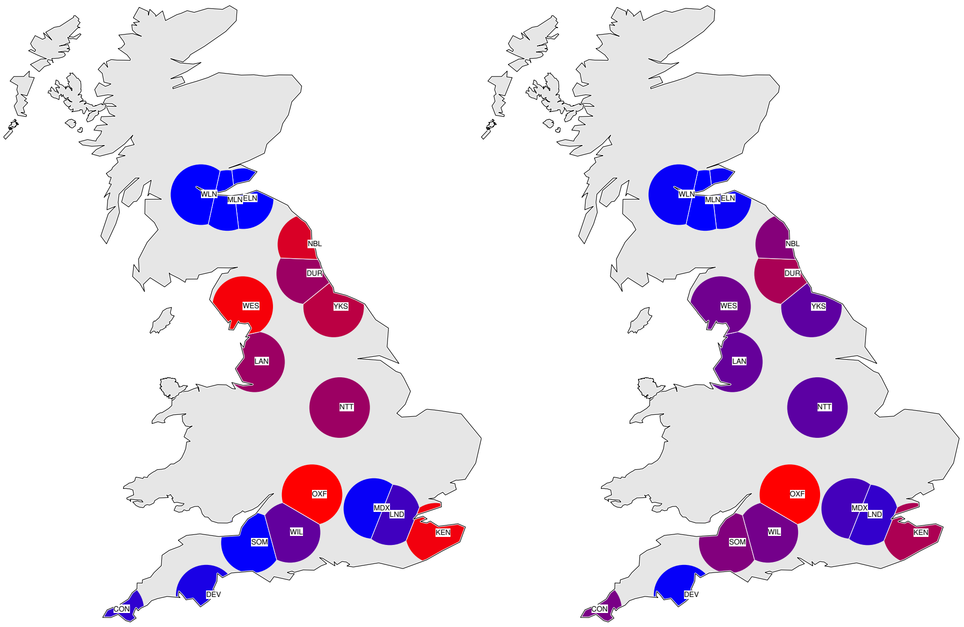
(46) -*nae* related bigrams

a. PPIS1.NN2
I_PPIS1 dinnae_NN2 like things like that at all [ELN_010]

b. PPIS2.NN2
We_PPIS2 coudnae_NN2 do it if [...] [ELN_008]

ngram	example	rel.	rank.rel.	n.sig.	rank.n.sig	N
PPIS1.NN2	*I havenae*	1.0	11	34	582	80
PPIS2.NN2	*we boys*	2.5	84	39	470	65
PPHS2.NN2	*they cloths*	4.4	411	24	870	38
PPHS1.NN2	*she wouldnae*	4.8	562	17	1183	29
PPY.NN2	*you loads*	5.4	793	19	1091	50
PPIO2.NN2	*us ups*	5.7	974	19	1087	79
PPHO2.NN2	*them berths*	5.8	1042	36	532	268
PPH1.NN2	*it arts*	7.0	1779	10	1661	27
PPHO1.NN2	*him robes*	10.5	4292	0	6215	10
PPGE.NN2	*ours nightfighters*	15.2	8065	0	6114	1

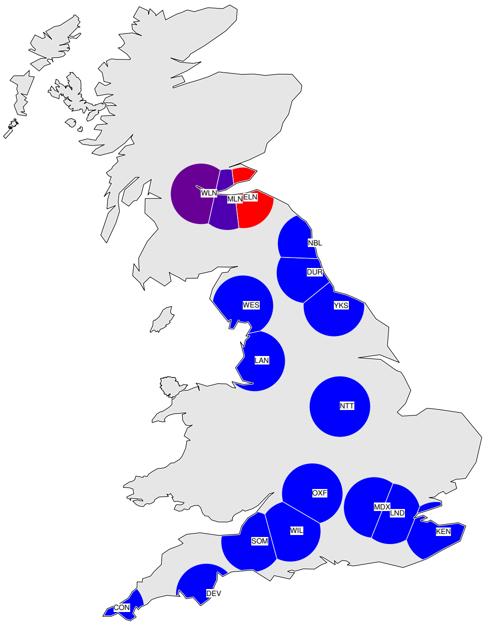
Table 4.17.: Most relevant bigrams related to *-nae*. Column *ngram* contains the POS bigram, column *example* an example. Column *rel* displays the reliability score (lower is more noteworthy), column *n.sig* the number of pairwise significant differences (higher is more noteworthy). Rank columns indicate the rank of this bigram in the total list when ordered by that metric. Final column shows total bigram frequency.



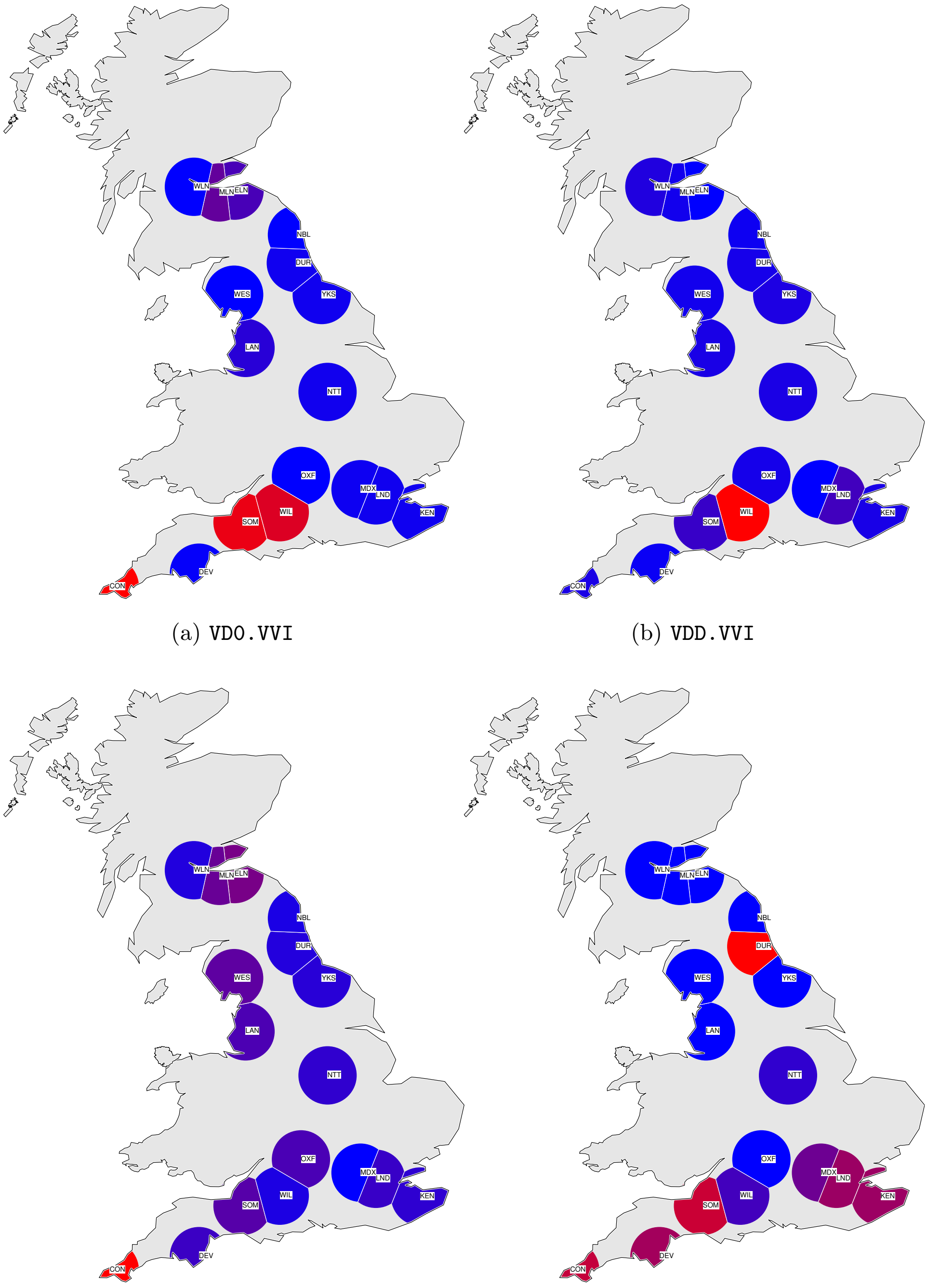
(a) `PPHO2.NNT2`

(b) `PPHO2.NN2`

Map 32: Visualization of geographic variation involving *them* + noun. More reddish colors indicate greater frequency of the bigram in question.



Map 33: Geographical variation involving -*nae*: `PPIS1.NN2`. More reddish colors indicate greater frequency of this bigram.



(a) `VD0.VVI`

(b) `VDD.VVI`

(c) `VD0.XX`

(d) `PPIS1.VDN`

Map 34: Visualization of geographic variation involving *do*. More reddish colors indicate greater frequency of the bigram in question.

c. `PPIS2.NN2`
So were we_`PPIS2` boys_`NN2` ; [. . .] [SOM_005]

(46a) and (46b) illustrate the patterns. Note that for plural pronouns, the same pattern may appear in contexts not involving -*nae*, such as in (46c). It is therefore not surprising that the `PPIS1.NN2` bigram emerges by far as the most distinctive according to reliability scores. Map 33 illustrates this: the `PPIS1.NN2` pattern occurs almost exclusively in Scotland.

This pattern is essentially a misclassification and thus an error. However, it can be considered a "happy accident": first, it captures actual dialectal variation, and second, it serves as an example for the power of this analysis method to identify interesting patterns.

This feature is also included in the model-based analyses, where it proved problematic due to its extreme geographic distribution. Still, Map 18a on page 108 agrees on the clearly Scottish nature of this feature.

4.2.2.5. *do*

Table 4.18 shows the top bigrams related to various forms of *do*. Two of the top patterns involve a form of *do* – either *do* (tag: `VD0`) or *did* (tag: `VDD`) followed by an infinitival verb form, as in (47a) and (47b). Maps 34a and 34b display their geographical distribution. Both patterns are especially frequent in the Southwest of England. This suggests that these bigrams are capturing variation related to *do* as a habitual or unstressed tense marker. These are classical dialect features of the Southwest, especially for invariant *do* as in the most distinctive `VD0.VVI` pattern (cf. Kortmann 2004c: 2.2 and 2.4)

(47) *do* related bigrams

a. `VD0.VVI`
Well he do_`VD0` make_`VVI` several different things . [SOM_009]

b. `VDD.VVI`
[. . .] we did_`VDD` call_`VVI` it arts . [SOM_002]

c. `iVD0.XX`
I do_`VD0` n't_`XX` suppose that hare saw us . [KEN_010]

d. `VD0.XX`
[. . .] where he do_`VD0` n't_`XX` tread on. [KEN_010]

e. `PPIS1.VDN`
I_`PPIS1` done_`VDN` the same thing ! [LND_003]

Another highly distinctive feature is the frequency of *do not* or *don't*, with the tag pattern `VD0.XX`, as illustrated by (47c) and (47d). Of course, this is generally a standard

ngram	example	rel.	rank.rel.	n.sig.	rank.n.sig	N
VD0.VVI	*do make*	2.0	39	46	364	139
VD0.XX	*do n't*	2.2	52	74	74	2362
VDD.VVI	*did spread*	2.6	91	49	316	437
PPIS1.VDN	*I done*	3.0	132	18	1131	54
TO.VDI	*to do*	3.8	258	31	663	1124
PPY.VDD	*you did*	3.8	280	59	205	379
NN1.VDD	*mother did*	3.9	284	58	216	291
PPHS2.VDN	*they done*	3.9	285	27	768	58
VBDR.VDG	*were doing*	4.0	323	18	1143	51
PPHS1.VDD	*she did*	4.1	334	57	230	511

Table 4.18.: Most relevant bigrams related to *-do*. Column *ngram* contains the POS bigram, column *example* an example. Column *rel* displays the reliability score (lower is more noteworthy), column *n.sig* the number of pairwise significant differences (higher is more noteworthy). Rank columns indicate the rank of this bigram in the total list when ordered by that metric. Final column shows total bigram frequency.

combination. Invariant *do* as the third person singular word form ((47d), see Features 40/41 in Section 4.1.1.8.2) is a dialectal feature that may lead to higher frequencies of the VD0.XX pattern. Map 34c displays the geographical distribution of this feature. It is especially frequent in Cornwall, and has intermediate to lower frequencies elsewhere. This does not fit the distribution of Features 40/41, which had higher frequencies throughout the South.

A final *do*-related pattern to be discussed here is PPIS1.VDN, the combination of the first person singular pronoun and *done*. This non-standard agreement pattern, illustrated in (47e), is clearly a very southern feature: As Map 34d illustrates, *I done* appears in all dialects in the Southeast and the Southwest of England except Oxfordshire, but not in any Northern English or Scottish dialect except for Durham. This result is very similar to that for non-standard *done*, Feature 29 (Section 4.1.1.6.3).

4.2.3. Sociolinguistic results: gender

This section discusses the n-grams where there is a significant gender difference in their distribution. As the gender split results in a binary distinction, reliability measures as discussed in Section 3.2.3 are not applicable, and the significance of the difference according to the method of Nerbonne & Wiersma (2006) will be used instead. A description of this can be found in that section as well. The discussion will proceed as follows: First, analysis will be restricted to those n-grams that exhibit a significant gender difference. These will be grouped by the type of tags they contain. For each n-gram, the normalized frequencies by male and female speakers will be provided, as well as the p-value resulting from the permutation test and whether that n-gram is preferred by male or female speakers. As there are generally a large number of n-grams belonging to a particular group, the lists will be usually be restricted to the n-grams with the largest absolute differences between male and female speakers. This restriction also removes very infrequent n-grams, which, even if significant, are particularly likely to be accidental.

Of the 221 tag unigrams[9], 48 (i.e. 22%) emerge as significant. Similarly, there are 8959 tag bigrams, and 643 of them (7%) show a significant gender difference in their distribution.

4.2.3.1. Nouns and Pronouns

Table 4.19 lists the tag unigrams related to determiners, nouns and pronouns that exhibit a significant gender difference in their distribution. A clear pattern emerges: all significant differences for nouns and articles are such that male speakers use them more often. Personal pronouns show the inverse pattern: with the exception of the third person plural pronoun *they* (PPHS2) all significantly different pronouns are used more often by female speakers. This matches the relevant results from the feature-based data set, where Feature 5, *us* showed a significant gender difference in the same direction. Determiners are, like nouns, consistently more often used by men, with the exception of possessive pronouns, which may function as a determiner and behave like most other pronouns, i.e. are more frequently used by women.

Based on the results of Feature 6, *them*, we would expect the unigram PPHO2 to appear in this list. And in fact, the gender difference between these features is about as large as that for *us*. However, it fails to achieve significance in the permutation test. This should not necessarily be seen as an indicator that the effect of gender in the model is spurious: there are fewer speakers in this data set, and the missing unigram significance may be the result of lower power. It is a useful reminder, however, that the modeling results should not be trusted without verification.

Table 4.20 shows bigrams with a significant and large gender difference involving one of the tags from Table 4.19 as the first constituent; bigrams where one of the relevant tags is the second component can be found in Tables 4.22 and 4.24. The patterns evidenced there mostly remain consistent: tag combinations involving articles and nouns are more often

[9] The number of n-gram types is slightly different from the one in the previous sections, as those texts in which gender information is not available are not included here.

ngram	example	m	f	p	by
APPGE	*mi*	2.08	2.93	0.000	f
AT	*the*	9.90	8.37	0.010	m
AT1	*a*	5.69	5.06	0.000	m
DD1	*that*	4.24	3.54	0.000	m
DDQ	*what*	1.28	1.09	0.046	m
NN1	*Bridge*	20.68	18.50	0.000	m
NNT2	*weeks*	0.88	0.72	0.028	m
NNU1	*ha'penny*	0.29	0.18	0.017	m
PN1	*One*	1.25	1.48	0.010	f
PPHS1	*she*	3.67	4.54	0.043	f
PPHS2	*they*	4.19	3.44	0.011	m
PPIO1	*me*	0.46	0.65	0.024	f
PPIO2	*us*	0.24	0.49	0.000	f
PPIS1	*I*	5.74	6.92	0.018	f
PPIS2	*we*	2.21	3.03	0.001	f

Table 4.19.: Unigrams relating to nouns and pronouns with a significant gender distribution. Column *ngram* contains the POS tag, column *example* an example. Columns *m* and *f* show the normalized frequencies for male and female speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the gender for which this unigram is most frequent.

ngram	example	m	f	p	by
APPGE.NN1	*my mother*	55.1	82.5	0.000	f
APPGE.NN2	*your ears*	12.6	18.6	0.000	f
AT1.NN1	*a Terrier*	135.1	119.5	0.003	m
AT.JJ	*the following*	51.9	38.1	0.002	m
AT.NN1	*the labour*	221.4	186.2	0.018	m
DD1.NN1	*another thing*	39.1	29.9	0.001	m
NN1.II	*rag on*	72.9	58.6	0.000	m
NN1.IO	*foot of*	66.0	50.0	0.000	m
NN1.NN1	*sell wash*	75.0	65.2	0.041	m
NN1.RL	*message home*	15.8	11.0	0.005	m
NN1.RP	*way round*	26.2	17.5	0.000	m
NN1.VV0	*school play*	15.8	7.5	0.000	m
PPH1.RP	*it back*	13.8	9.4	0.005	m
PPH1.VBDZ	*it was*	39.3	48.8	0.027	f
PPHO2.RP	*them up*	9.4	5.3	0.000	m
PPHS1.VBDZ	*he was*	22.2	31.7	0.013	f
PPHS1.VM	*he would*	14.8	20.5	0.006	f
PPHS2.VHD	*they had*	17.2	12.0	0.002	m
PPHS2.VM	*they could*	25.3	14.8	0.001	m
PPHS2.VV0	*they work*	14.5	10.5	0.003	m
PPIS1.RR	*I often*	7.3	12.8	0.000	f
PPIS1.VBM	*I 'm*	6.6	10.6	0.022	f
PPIS1.VM	*I ca*	25.0	37.3	0.000	f
PPIS2.VBDR	*We were*	5.4	10.5	0.006	f
PPIS2.VHD	*we 'd*	13.3	18.9	0.028	f

Table 4.20.: Bigrams relating to nouns and pronouns with a significant gender distribution and large gender difference, first component. Column *ngram* contains the POS bigram, column *example* an example. Columns *m* and *f* show the normalized frequencies for male and female speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the gender for which this bigram is most frequent.

ngram	example	m	f	p	by
VDD	*did*	0.84	1.12	0.001	f
VDI	*do*	0.37	0.48	0.020	f

Table 4.21.: Unigrams relating to verbs with a significant gender distribution. Column *ngram* contains the POS tag, column *example* an example. Columns *m* and *f* show the normalized frequencies for male and female speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the gender for which this unigram is most frequent.

used by male speakers, combinations involving pronouns (including possessive pronouns) are more often used by female speakers. Regarding pronouns, there are few exceptions to the pattern evidenced in Table 4.19. Combinations involving the third person plural subject pronoun are still much more frequent for male speakers, in contrast to to their general preference for lexical nouns over pronouns. Tag combinations of pronouns and particles (RP) also tend to be preferably used by men, which is likely to be a side effect of men using these words more.

4.2.3.2. Verbs

Table 4.21 displays the tag unigrams involving verb forms that show a significant gender distribution and large gender difference. This list is very short, containing only two tokens involving *to do* that are used more frequently by female speakers. Table 4.22, displaying bigrams beginning with a verb form, expands on this: several forms of *to be*, *to do* as well as modal verbs are used more often by women. Lexical verbs, on the other hand, appear mostly in patterns strongly favored by men. This pattern does not hold when considering the patterns that have a relatively small gender difference, and thus indicates that this list may result largely from the other word: the negator (XX) and, as with pronouns, prepositional adverbs and particles RP. For *to do*, however, the pattern does hold: only the finite base form and *done* appear in a bigram used more by men. This dovetails nicely with the gender effect found in the models for Feature 13, *to do*.

However, as with *them*, we would expect a difference for *used to* (VMK.TO), which the models for Feature 19/20 identified as more characteristic for female speakers. The difference is only present in absolute numbers, but it is not significant.

ngram	example	m	f	p	by
VBDZ.JJ	*was late*	12.5	18	0.000	f
VBDZ.RR	*was just*	13.1	17	0.003	f
VBDZ.XX	*was n't*	9.1	14	0.000	f
VDD.XX	*did n't*	19.4	27	0.001	f
VM.XX	*would n't*	25.5	33	0.002	f
VV0.AT	*know the*	17.1	13	0.001	m
VV0.RP	*get up*	24.7	17	0.000	m
VVD.RP	*went down*	34.1	27	0.006	m
VVI.II	*count to*	28.7	36	0.000	f
VVN.RP	*Swollen up*	16.2	11	0.001	m

Table 4.22.: Bigrams relating to verbs with a significant gender distribution and large gender difference, first component. Column *ngram* contains the POS bigram, column *example* an example. Columns *m* and *f* show the normalized frequencies for male and female speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the gender for which this bigram is most frequent.

4.2.3.3. Other

Table 4.23 displays the remaining tag unigrams with a significant and at least somewhat large gender difference. As was already noted, female speakers use the negator *not/n't* (XX) more often, and the same is true for adverbs (RR) and interjections (UH). Men, on the other hand, exhibit greater frequencies of prepositions, prepositional adverbs and particles. Table 4.24, containing bigrams that have one of the tags from Table 4.23 as their first constituent, and the relevant parts of Tables 4.20 and 4.22 largely confirm these patterns. One thing emerges that is not clear from the unigram-based table: Several patterns with a coordinating conjunction are used more often by women and only one is used more by men. As with lexical verbs, this appears to be an issue of concentration: the patterns with small differences are those used more often by men.

4.2.3.4. Interim summary

To summarize, the bottom-up analysis has uncovered many significant differences between male and female speakers in FRED-S. One of the major differences is that, on the whole, female speakers use more pronouns, while male speakers use more lexical nouns; similarly female speakers use more primary verbs – especially *to do* – interjections, and negators, while male speakers use more base forms of lexical verbs.

ngram	example	m	f	p	by
CSA	*as*	0.53	0.36	0.009	m
II	*in*	11.15	9.80	0.000	m
IO	*of*	2.66	2.05	0.000	m
MC	*two*	2.97	2.43	0.007	m
RP	*down*	5.57	4.07	0.000	m
RR	*ever*	6.31	7.32	0.002	f
UH	*yes*	6.88	9.24	0.005	f
XX	*n't*	2.45	3.09	0.002	f

Table 4.23.: Other unigrams with a significant gender distribution and large gender difference. Column *ngram* contains the POS tag, column *example* an example. Columns *m* and *f* show the normalized frequencies for male and female speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the gender for which this unigram is most frequent.

Let us now compare these results to those for the model-based analyses in Section 4.1.1, especially Table 4.4 (page 149). Some of the significant differences in that list can in some way be found in the bigrams, albeit as relatively rare features that therefore have an overall low gender difference and may not appear in the tables. Consider Feature 49 *as what* or *than what* in comparative clauses: the bigram counts for these strings are captured by CSA.DDQ and CSN.DDQ. Both have a significant or marginally significant gender distribution, with male speakers having about twice the normalized frequency that female speakers do. Feature 5, *us*, and Feature 13, *to do*, were already discussed above.

That said, some patterns that were expected did not achieve statistical significance: *them* (Feature 6) and *used to* over *would* (Features 19/20). This could result from missing power, as the bottom-up analysis is based on a smaller corpus. Nevertheless, it should serve as a warning sign against taking either modeling or n-gram results too literally. Where the results are robust and in concord with the existing literature, such as pronouns matching the results from Rayson et al. (1997) and Hirschman (1994), they can serve as additional support.

In total then, what models and bottom-up analyses tell us about the effects of gender is partially similar. The next section will test whether the same is true for speaker age.

4.2.4. Sociolinguistic results: age

This section discusses the n-grams where there is a significant age difference in their distribution. The method and presentation closely follows the structure outlined in Section

ngram	example	m	f	p	by
CC.APPGE	*and our*	4.7	9.6	0.000	f
CCB.PPIS1	*but I*	6.5	11.2	0.000	f
CC.MC	*or six*	17.2	10.9	0.000	m
CC.PPHS1	*and he*	22.2	30.9	0.016	f
CC.PPIS1	*And I*	21.2	30.0	0.003	f
CC.PPIS2	*And we*	9.4	15.2	0.009	f
CC.RT	*and then*	25.7	35.0	0.031	f
CS.PPHS1	*'cause he*	10.3	16.9	0.000	f
CS.PPIS1	*when I*	19.6	26.9	0.008	f
II21.II22	*on to*	20.6	15.8	0.032	m
II.AT	*in the*	147.8	124.1	0.016	m
II.AT1	*in a*	30.4	25.5	0.026	m
II.II	*about in*	10.5	5.2	0.000	m
II.RL	*on there*	15.2	10.1	0.010	m
IO.AT	*of the*	22.0	13.8	0.000	m
JJ.NN1	*strong man*	105.6	90.0	0.011	m
RG.JJ	*very untidy*	17.2	23.1	0.005	f
RP.AT	*up the*	10.0	5.2	0.000	m
RP.II	*up through*	55.1	35.6	0.000	m
RP.RL	*up there*	26.7	17.4	0.011	m
RR.PPIS1	*So I*	18.9	25.9	0.009	f
RR.UH	*well ah*	6.3	12.1	0.006	f
RR.VVD	*always said*	12.5	20.4	0.000	f
UH.PPH1	*aye it*	7.3	12.0	0.000	f
UH.PPHS1	*Phew he*	10.8	19.4	0.000	f
UH.PPIS1	*Oh I*	18.1	31.6	0.000	f
UH.PPIS2	*Eh we*	7.6	14.5	0.000	f
UH.RR	*Oh definitely*	12.9	17.9	0.009	f
XX.VVI	*n't believe*	46.9	62.2	0.001	f

Table 4.24.: Bigrams containing other tags with a significant gender distribution and large gender difference. Column *ngram* contains the POS bigram, column *example* an example. Columns *m* and *f* show the normalized frequencies for male and female speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the gender for which this bigram is most frequent.

4.2.3. The speakers were mapped into two age groups, and this mapping attempted to create groups that contain similar numbers of speakers while keeping the ratio of male to female speakers in both subcorpora as equal as possible. The optimal cut-off point was determined to lie at a speaker age of 80 years. Unfortunately there still remains a gender bias such that younger speakers are more likely to be female than older speakers are.

Of the 221 tag unigrams, 19 (i.e. 9%) emerge as significant. Similarly, there are 8748 tag bigrams, and 308 of them (4%) show a significant age difference in their distribution.

4.2.4.1. Nouns and pronouns

Table 4.25 displays the tag unigrams related to determiners, nouns and pronouns that exhibit a large and significant difference between age groups. There is a clear similarity to the gender effect for the same tags, with younger speakers, like women, tending to use more pronouns, and older speakers, like men, using more determiners, lexical nouns, and third person plural pronouns (here including *them*). Table 4.26 shows the relevant bigrams involving one of these tags as the first constituent, with Tables 4.28 and 4.30 containing the same as the second constituent. As usual, the unigram pattern is largely reflected in the bigrams, with a few exceptions: singular determiners followed by an adjective (`AT1.JJ`) are more frequently used by younger speakers, although singular determiners (`AT1`) and nouns (`NN1`) are in general more frequently used by older speakers. One bigram that should be highlighted is *they was*, `PPHS2.VBDZ`. Its greater use by older speakers neatly matches the effect of age found in the models for Feature 44, non-standard *was*.

4.2.4.2. Verbs

Table 4.27 lists the small number of tag unigrams related to verb forms that exhibit a large and significant difference between age groups. Whereas the previous section found a lot of similarities between gender and age differences, there are none to be found here. We find two forms of present tense *to be* that are used more by male speakers, the marker of past habituality *used to*.

Table 4.28 shows the bigrams where a verb form is the first constituent that have a significant and relevant age-related distribution; bigrams with the verb as the second constituent can be found in Tables 4.26 and 4.30. As expected from the unigram distribution, *used to* (`VMK.TO`) is used more often by older speakers, and this difference is very large compared to the other differences between age groups. This confirms the results of the model-based analyses for Feature 19/20, *used to/would* (Section 4.1.1.4.2), for age.

ngram	example	old	young	p	by
APPGE	*his*	2.30	2.68	0.000	young
AT	*the*	9.33	8.32	0.000	old
AT1	*a*	5.39	5.19	0.023	old
DD	*some*	0.66	0.54	0.000	old
NN	*people*	0.64	0.53	0.001	old
NN2	*prostitutes*	5.59	5.37	0.029	old
NNT1	*night*	1.91	1.66	0.000	old
PPH1	*it*	3.87	4.22	0.000	young
PPHO1	*him*	0.57	0.71	0.000	young
PPHO2	*them*	1.69	1.27	0.000	old
PPHS1	*she*	3.67	4.45	0.000	young
PPHS2	*they*	4.08	3.35	0.000	old
PPIO1	*me*	0.49	0.63	0.000	young
PPIS1	*I*	5.92	6.43	0.001	young
PPIS2	*we*	2.46	2.75	0.000	young
PPY	*you*	4.86	5.61	0.000	young

Table 4.25.: Unigrams relating to nouns and pronouns with a significant age distribution. Column *ngram* contains the POS tag, column *example* an example. Columns *young* and *old* show the normalized frequencies for younger and older speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the age group for which this unigram is most frequent.

ngram	example	young	old	p	by
AT1.JJ	*a gross*	45.4	38.7	0.007	young
PPH1.VBZ	*It 's*	19.9	12.6	0.001	young
PPHS1.VBZ	*He 's*	7.6	3.4	0.000	young
PPHS2.VBDZ	*they was*	3.9	11.0	0.006	old
PPHS2.VHD	*they had*	12.1	16.4	0.005	old
PPHS2.VMK	*they used*	13.3	26.5	0.001	old

Table 4.26.: Bigrams relating to nouns and pronouns with a significant age distribution and large age difference, first component. Column *ngram* contains the POS bigram, column *example* an example. Columns *young* and *old* show the normalized frequencies for younger and older speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the age group for which this bigram is most frequent.

ngram	example	young	old	p	by
VBR	*'re*	0.51	0.33	0.002	young
VBZ	*is*	2.28	1.68	0.003	young
VMK	*used*	2.09	3.05	0.036	old

Table 4.27.: Unigrams relating to verbs with a significant age distribution. Column *ngram* contains the POS tag, column *example* an example. Columns *young* and *old* show the normalized frequencies for younger and older speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the age group for which this unigram is most frequent.

ngram	example	young	old	p	by
VBZ.JJ	*'s true*	13	7.2	0.025	young
VMK.TO	*used to*	88	128.2	0.032	old
VVI.AT	*pay the*	16	20.6	0.036	old
VVI.PPHO2	*afford them*	11	17.8	0.001	old

Table 4.28.: Bigrams relating to verbs with a significant age distribution and large difference between age groups, first component. Column *ngram* contains the POS bigram, column *example* an example. Columns *young* and *old* show the normalized frequencies for younger and older speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the age group for which this bigram is most frequent.

ngram	example	young	old	p	by
JJ	*old*	6.56	5.89	0.021	young
TO	*to*	4.22	5.48	0.007	old

Table 4.29.: Other unigrams with a significant age distribution and large age difference. Column *ngram* contains the POS tag, column *example* an example. Columns *young* and *old* show the normalized frequencies for younger and older speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the age group for which this unigram is most frequent.

ngram	example	young	old	p	by
CC.VV0	*and work*	20.5	28	0.022	old
CS.PPHS2	*when they*	14.1	18	0.023	old
II.PPHO2	*in them*	8.3	13	0.045	old
TO.VHI	*to have*	12.3	20	0.025	old
TO.VVI	*to say*	127.8	160	0.011	old

Table 4.30.: Bigrams containing other tags with a significant age distribution and large difference between age groups, first component. Column *ngram* contains the POS bigram, column *example* an example. Columns *young* and *old* show the normalized frequencies for younger and older speakers. Column *p* displays the permutation-based significance of the difference. Final column indicates the age group for which this bigram is most frequent.

4.2.4.3. Other

Table 4.29 displays the tag unigrams that exhibit a large and significant difference between age groups and were not included in Tables 4.25 and 4.27; again this list is very short. The few noteworthy differences have already appeared in the previous discussions, namely the greater frequency of adjectives in the material from younger speakers or of TO, often as part of an infinitive, by older speakers. Table 4.30 does not add a lot, but confirms the *used to* pattern once again.

4.3. Chapter summary

This chapter began with the complete results of the lmer models and GAMs. It was found that the features did exhibit a spatial distribution in most cases, and this distribution usually fit together neatly with previous dialectological research on these features. Then a case study explored the effects of more elaborate investigation. This analysis was found to be essentially similar to the simple models, as evidenced by a high correlation coefficient between the geographic results. Crucially, the simple models matched the complex models better than the normalization-based values did. The sociolinguistic summary confirmed the hypotheses: female speakers use fewer non-standard features and older speakers use them more frequently. In the geographic summary, features with a noteworthy spatial distribution were presented, and it was shown that the features can be divided into four groups: Southern English features, Southeastern English features, Scottish features and Highlands/Hebrides features.

The bottom-up analysis was able to identify many geolinguistically interesting patterns. Most of them were already included in the feature set used for modeling, yet the bottom-up approach yields frequency information about more narrow contexts, which may add to the overall result. The sociolinguistic analysis, using speaker-based permutation, resulted in a selection of uni- and bigrams that were significantly different in their gender or age distribution across the FRED-S corpus. These partially reflected the effects observed in the models, and did not contradict them. On the other hand, some differences that one might expect on the basis of the previous analyses were not found, or did not achieve statistical significance.

The next chapter will present the results of aggregate analysis on both modeling and bottom-up results. How good these results are will be evaluated in the final chapter, using different operationalizations of the distance between locations as a yardstick.

5. Aggregational analyses

This chapter reports the results of using dialectometric methods on the results of the lmer models, GAMs, and n-gram analyses discussed in the previous chapter. A discussion of how the three types of top-down analysis compare to one another will be provided first. Then, hierarchical cluster analyses of the results of these three methods will be provided. This will be followed by similar cluster maps for four variants of bottom-up analyses: frequencies and reliability scores for both unigrams and bigrams. Finally, three of the distance matrices - those based on lmer models, GAMs and bigram reliability scores - will be subjected to analysis with NeighborNet and continuum maps.

Let me briefly summarize the methodology used to derive distances from the models or n-grams and, ultimately, from feature frequencies. A more detailed description can be found in Chapter 3. The original version of CBDM presented in Szmrecsanyi (2013) can be summarized with the following 'cooking recipe':

- count the frequencies of the features under study in the dialect corpus
- normalize the frequencies to make the areas comparable
- use a logarithmic transformation to de-emphasize the influence of overall frequency, set a lower limit to -1
- derive a distance matrix using an appropriate distance function, here the Euclidean distance

My analysis largely follows the same steps, but replaces the second step with the following:

- create a model for the number of uses of the feature, based on the available sociolinguistic factors, and one of the following operationalizations of geography:
  - a categorical factor, specifying the county that a speaker is from, used as a *random effect*. This leverages the partial pooling effect, moving cases toward the grand mean inversely proportional to the strength of the evidence for that group

  - as longitude and latitude of individual interview sites, used as a thin plate regression spline in a generalized additive model. This pulls the values for individual sites, in the absence of strong evidence, toward those of their neighbors

- use that model to predict how many tokens of that feature a speaker from a certain county would use in 10.000 words

- proceed as above, using model predictions instead of the normalized values.

For n-grams, normalized frequencies were counted, and the reliability scores were determined using a permutation-based process. The corpus was resampled based on conversational turns, and the new random subcorpora were compared to the original data set based on normalized n-gram counts. Particularly low reliability scores indicate reliably high frequencies, particularly high reliability scores indicate reliably low frequencies.

All cluster analyses and maps in this section were created using Peter Kleiweg's dialectometry software package *RuG/L04*.

5.1. Comparing normalization- and model-based approaches

Let us begin by comparing the two model-based approaches to the normalization-based method and to each other. The reasoning in Section 3.2.2 would lead to the following hypotheses: compared to normalization, mixed-effect modeling should lead to reduced distances for counties that do not have a large amount of data available, while the GAM-based method should lead to lower distances for counties that are geographically close.

The evaluation of these hypotheses is not quite straightforward, as the absolute difference in distances between methods alone is not necessarily informative. Both model variants should, on average, bring extreme values closer to the mean, and therefore the distances resulting from their predictions should be lower across the board. I therefore turn to regression modeling of the distances. One distance is used as the predictor, and the regression model estimates the effect that an increase in that distance has on distances resulting from a different method. Figure 5.1 illustrates this, plotting lmer-derived distances (on the y-axis) against normalization-based distances (on the x-axis). The black line indicates the overall relationship: as one set of distances increases, the other increases as well. The relationship is not perfect, and individual distances diverge from this pattern to varying degrees. The further a pair of distances is from the main trend, the greater the difference in how the two methods evaluate the pair of these two counties. The points above the line are distances that the lmer model emphasizes, whereas the points below the line are distances that the normalization-based method emphasizes. Overall, the

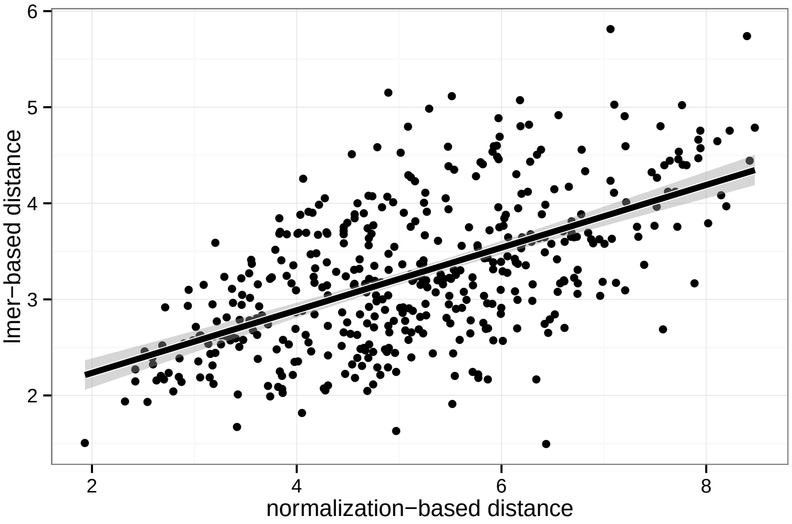
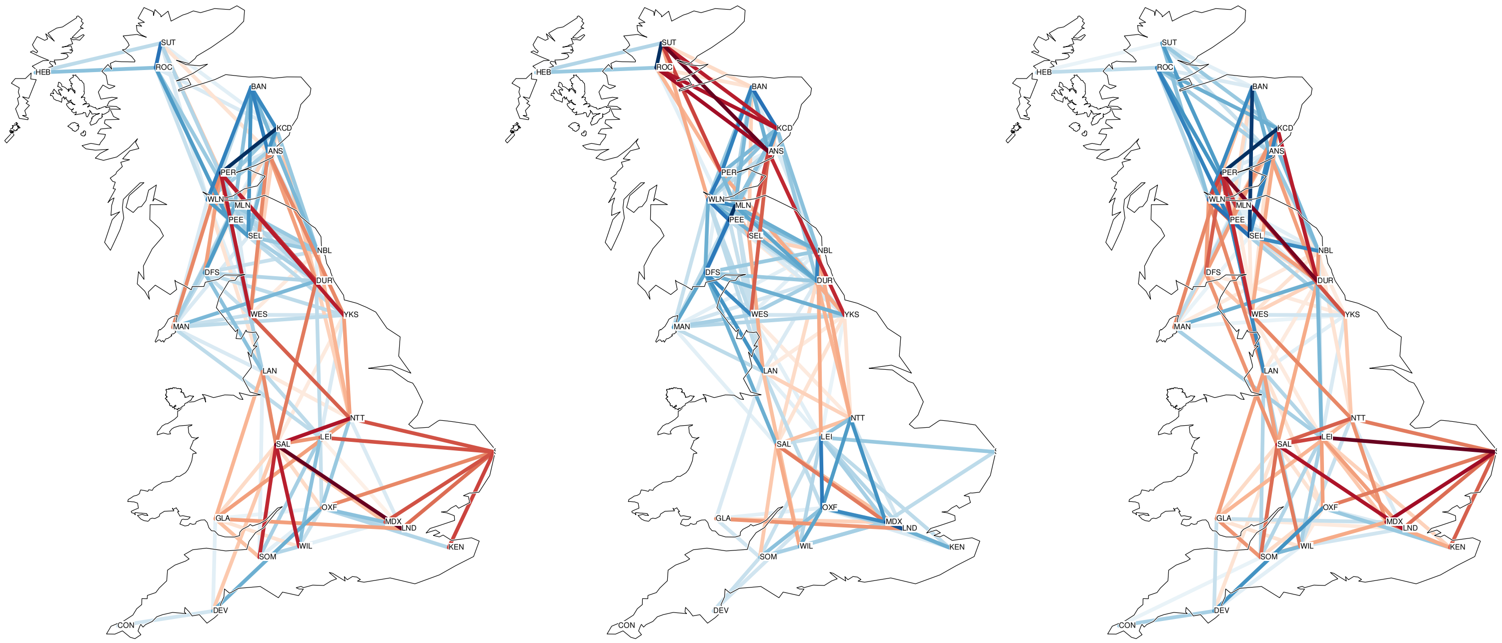


Figure 5.1.: Distances resulting from normalization compared to distances resulting from lmer modeling. The black line indicates the average relation between both distances. Points above the line are distances the lmer model emphasizes, points below the line are distances the normalization-based method emphasizes.

normalization-based distances explain 32 percent of the variability in the lmer-derived distances and 42 percent in the GAM-based distances. The mapping between the two models is even better at 50 percent of the variance.

Map 35 displays the result for all pairwise combinations of methods as line maps. For interpretatory convenience, lines between pairs that are particularly far apart are not shown. In all maps, the more blue a line is, the more the first method considers the difference between the two points to be larger than the second, while more red lines indicate that the second method considers the difference to be larger than the first. In other words, the further a given distance is above the line in Figure 5.1, the deeper the shade of red in Map 35a, and the further it is below that line, the deeper the shade of blue.

For the comparison of normalized results to lmer model results, shown in Map 35a, we find that most distances in Scotland are colored blue, i.e. the distances resulting from lmer model predictions are smaller. This is exactly as expected, considering that the coverage in running words for many counties is rather low in Scotland (cf. Map 2a on page 45). On the other hand, distances involving the counties with good coverage, particularly Shropshire and Suffolk, tend to be higher. The same is true for distances between broad areas, such as those between the North of England and the Scottish Lowlands, or between



(a) normalized and lmer distances($R^2 = 0.32$) (b) normalized and GAM distances ($R^2 = 0.42$) (c) GAM and lmer distances ($R^2 = 0.50$)

Map 35: Comparison of methods: differences between normalization-based and lmer and GAM derived distances as resulting from a linear regression model. Red lines indicate the second method leads to greater distances, blue lines indicate the first method leads to greater distances.

the North and the Midlands, with the exception of Leicestershire.

The comparison of GAM and normalized values in Map 35b leads to a picture with some crucial differences. First, the GAM-based method finds much stronger support for separating the Scottish Lowlands from the Highlands, which show closer connections to each other and to the Hebrides than they do in the lmer model predictions. Again, the Scottish Lowlands form a more cohesive group. The difference between the Lowlands and the English North is less pronounced than in the previous map; in general, the GAMs predict this difference to be smaller than the normalized frequencies. Exceptions to this exist, however, in particular the distances between Angus and many other counties in the Lowlands and the North of England. In England, the strong differences involving Suffolk and Shropshire that was observed in the previous map almost completely vanishes. Shropshire, however, is still markedly more distant from the Southern dialects than in the normalized distances. The rest of the South generally forms a more cohesive group here, while its distance to the North of England is slightly higher.

Map 35c, finally, compares both model types against one another. Here, red lines indicate smaller distances for GAMs, and blue lines indicate higher distances for lmer models. Clearly, the GAM method places more emphasis on the difference between the Scottish High- and Lowlands as well as generally on higher geographic distances, whereas the lmer model weighs the differences between the counties with particularly good coverage more.

5.2. Dialect areas: hierarchical clustering

This section concerns itself with the classification, more precisely the hierarchical grouping, of dialects. To do so, I use a hierarchical clustering algorithm. In essence, such analysis moves upward from the individual points, finding those with the smallest distance, then fusing them into a single unit. The distances between the new unit and the other points are then recalculated, and the process is iterated on the resulting matrix until only a single point remains. The order in which the elements were merged can then be interpreted as a classification. For example, to divide all the points into two groups, the analyst looks back to the point immediately before the last fusion. At that point, the data set consisted of two points, each of which may represent many points. This splits the set of points into two parts, and both can be considered a group. An example of how such an analysis proceeds can be found in Section 2.1.2.

There are a lot of parameters that the researcher can change while operating a cluster analysis. First, there are many methods that can be used to choose which points to fuse, and how the distance to the new unit should be calculated. For the following, I always use "Ward's method", a common choice in dialectometry (Goebl 2006, Sanders 2010), yet not an uncontroversial one (see Heeringa 2004). The reason is that it is the algorithm chosen by Szmrecsanyi (2011), and one of the two algorithms used in Szmrecsanyi (2013). As one of the goals of the present investigation is comparability with these studies, introducing additional variation at this point is unnecessary. The second choice concerns the number of clusters that are considered for evaluation, especially for plotting. The usual procedure is an inductive and somewhat subjective choice based on how explanatory the clusters are, using a measure called the fusion coefficient and scree plots (see Szmrecsanyi 2013: 118). Here, I keep this number fixed at 5, the number of clusters used in Szmrecsanyi (2011). Again this is motivated by the goal of maximizing the comparability to that study. Szmrecsanyi (2013) determines three clusters as the optimal value, but this hides some of the discontinuities in the result. The number of clusters here should be seen mostly as a visual aid; the full classification structure can always be found in the dendrograms that accompany the maps.

Clustering is a process where small changes to the data can have large effects. To mitigate this problem, I use a method proposed by Nerbonne et al. (2008), who suggest adding small amounts of random numeric noise to the distance matrix before clustering, then repeating this process a large number of times. The results of this process can then be aggregated into a new distance matrix. I use a noise setting of half the standard deviation of the distances, and repeat the process 10,000 times. The resulting distance matrix should be relatively robust to minor changes in the data.

The rest of this section proceeds as follows: First, I will revisit the original, unmodeled data used by Szmrecsanyi (2011; 2013). Only subsets of the data are suitable for the various analyses presented here; therefore it is necessary to establish first how normalized counts fare on these subsets. Then, the results of the models will be presented, first for lmer models and then for GAMs. Afterward, the distances resulting from unigram and bigram frequencies will be analyzed. Finally, I will present the results of distances based on a permutation-based measure, again for both unigrams and bigrams.

5.2.1. Normalization-based results

Before discussing the results of the new methodologies presented here, I first present the results of using Szmrecsanyi's CBDM on the reduced data sets. Due to the unavailability of information regarding speaker gender and age, only 273 of the 350 speakers (78 percent) in Szmrecsanyi (2013)'s original study are included here; considering the number of words, about 90 percent of the original 2,400,000 remain. This reduction also leads to the complete absence of three of the original 34 counties: East Lothian in Scotland, Denbighshire in Wales and Warwickshire in the English Midlands. How do these changes affect the output of Szmrecsanyi's method? The new results will serve as a baseline for the evaluation of the model-based and bottom-up analyses.

Map 36 displays the result of noisy hierarchical clustering using Ward's method. The main split in the data is broadly between, on the one hand, England, Wales, and the Isle of Man, which form the light and dark blue groups, and Scotland, represented as pink, red and green groups. Deviations from this larger pattern include Northumberland, which is usually considered to be part of the North of England but is grouped with Scotland here, and the Hebrides, which are part of the (mostly northern) English cluster. In slightly greater detail, England divides into two clusters, a light and a dark blue one. The first comprises all Northern English dialects except Northumberland, as well as Glamorganshire in Wales, the Isle of Man, Shropshire and Leicestershire in the English Midlands, Middlesex in the Southeast, and, as noted above, the Hebrides. The remaining English dialects, i.e. all Southeastern and Southwestern English dialects except for Middlesex, plus Nottinghamshire in the Midlands, form the second group. The Scottish dialects fall into three groups, a major one (in red) spanning most dialects in the Scottish Lowlands, a minor one (in green) containing Northumberland in the North of England as well as two dialects close to the English/Scottish border, Dumfriesshire and Peeblesshire, and finally, a pink cluster containing Midlothian and the Scottish Highlands.

Comparing these results to the five groups reported in Szmrecsanyi (2011), we can see that the overall results are quite similar, as was to be expected. The Southern English group is mostly unchanged here, with the exceptions of Durham in the Northeast, which joins that group in Szmrecsanyi (2011). The major Northern English group is similar as well, but now includes Durham, Middlesex, and the Hebrides. The division in Scotland is somewhat different between the two data sets. The split into a group of dialects close to the border and a main Lowlands group cannot be found there; instead, all these dialects form one group with the Lothians as a distinct sub-cluster. The final original cluster spans the Scottish Highlands, the Hebrides, as well as the British and Welsh outliers Denbighshire, Warwickshire and Middlesex. This group is the least similar to any in the reduced data set, presumably due to the fact that this cluster was affected the most by the removal of speakers. Its closest analogue is the red group comprising the Scottish Lowlands and Midlothian.

One customary way to examine how well linguistic and geographic distances fit is to correlate them using the Pearson product-moment correlation coefficient; for the reduced data set using Szmrecsanyi's method this statistic is 0.22, and geography explains 4.9

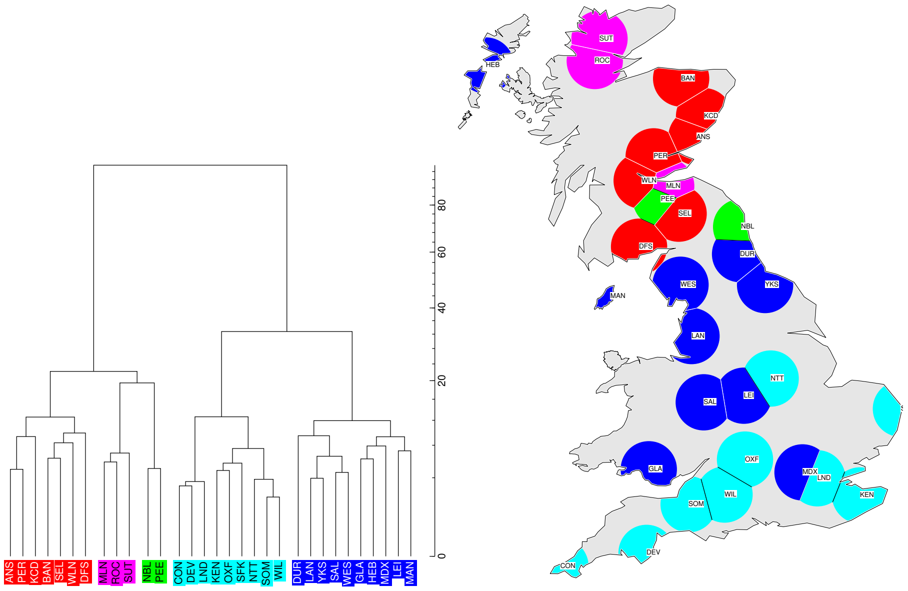
percent of the variance observed in the data. This is somewhat higher than the correlation for the full data set (0.21 and 4.4 percent), presumably due to the removal of two atypical counties. Furthermore, the groups emerging from the clustering process are geographically more contiguous, although outliers still remain.

For the bottom-up analyses, yet another subset of the data was used, as only the texts from the FRED-S subcorpus were available in a part-of-speech tagged version. Map 37 displays the result of Szmrecsanyi's strategy when analysis is restricted to that subset. The topmost split separates the three Scottish varieties in FRED-S (blue) from the English dialects. Most of the Southern English varieties group in a single light blue cluster, with the exception of Somerset and Wiltshire, which together with Nottinghamshire form the red cluster. The North of England, finally, falls into two groups: the green group, consisting of two English dialects close to the Scottish border (Northumberland and Westmorland), and the pink one spanning the remaining counties. This grouping seems geographically quite contiguous, which the correlation between geographic and linguistic distances confirms: $r = 0.52$, with 27.6 percent of the variance are explained.

5.2.2. lmer-based results

I now turn to the distance matrix resulting from observed frequencies processed using lmer models with Poisson regression and county as a random effect. Map 38 shows the result of noisy hierarchical clustering using Ward's method. Again, we find that the major split in the data is between England, represented by the red, light blue and dark blue clusters, and Scotland, comprising the pink and green clusters. The cluster boundaries, however, seem to fit the geographic pattern less well than for Szmrecsanyi's method. The major Southern English group, in red, now includes both Middlesex, which was part of the Northern cluster in Map 36, but no longer contains Somerset and Wiltshire (as, I note, holds when considering only texts in FRED-S). The group containing the British North, in dark blue, includes the Scottish Highlands now, but no longer Lancashire nor Leicestershire. The four counties missing from these groups form part of a new, geographically spread out cluster in light blue, which also contains two members of the previous English/Scottish border group, Dumfriesshire and Peeblesshire. Northumberland again forms part of a small group away from the Scottish main group, this time together with Midlothian. The northern Scottish Lowlands remain unchanged.

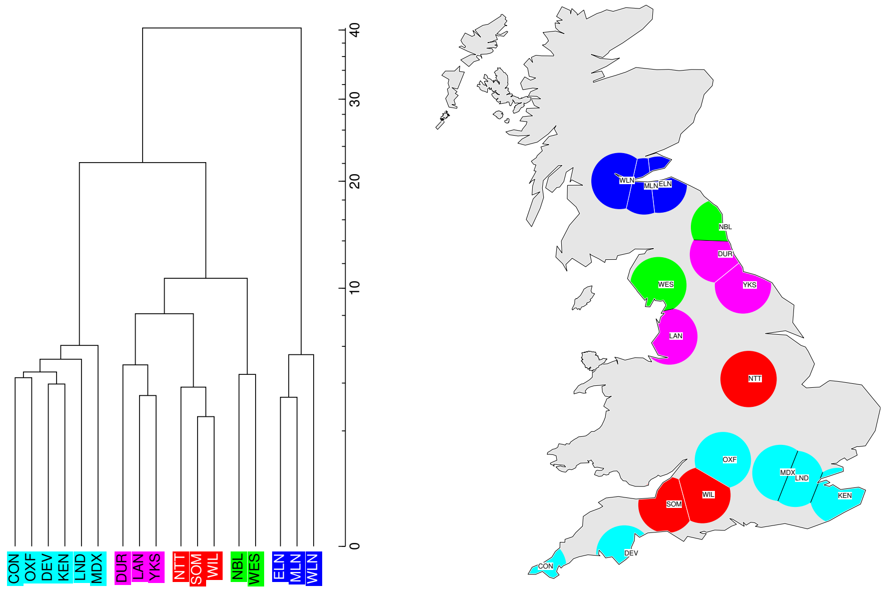
Overall, the picture resulting from the lmer models is similar to the one using normalization on a large scale, yet has notable differences in the details. Qualitatively, it is difficult to evaluate them - neither is geographically continuous, nor do the outliers match the previous classifications particularly well. Quantitatively, however, we can again



(a) Dendrogram

(b) Map

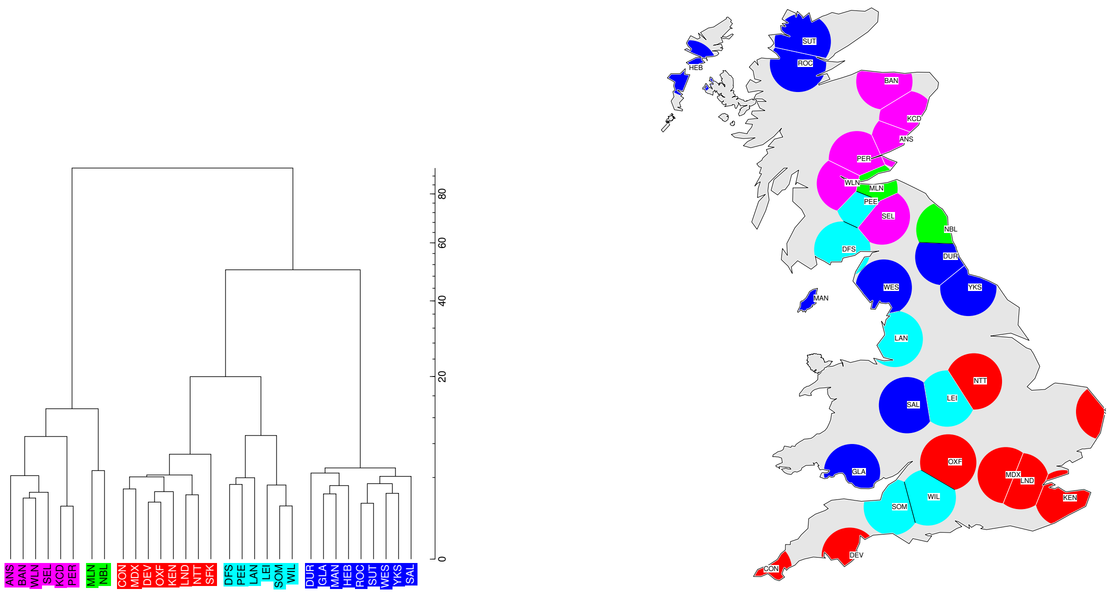
Map 36: Cluster analysis based on normalized feature frequencies. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.



(a) Dendrogram

(b) Map

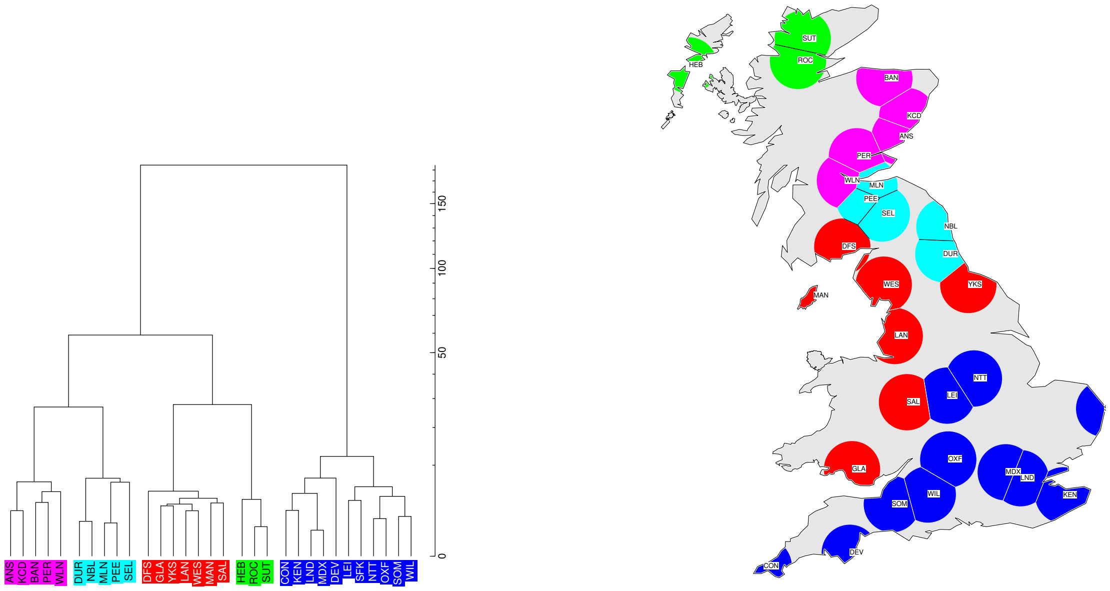
Map 37: Cluster analysis based on normalized feature frequencies, only texts in FRED-S. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.



(a) Dendrogram

(b) Map

Map 38: Cluster analysis based on lmer model. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.



(a) Dendrogram

(b) Map

Map 39: Cluster analysis based on GAM. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.

compare the match between linguistic and geographic distances. Doing so results in a correlation coefficient of $r = 0.32$, explaining 10.1 percent of the variance. Furthermore, the relationship between linguistic and geographic distances seems to be sublinear; comparing linguistic distances with logarithmically transformed geographic distances, a correlation coefficient of $r = 0.35$ is achieved, accounting for 12 percent of the variance. This, in contrast to what was observed for the full data set by Szmrecsanyi and for the reduced data set above, fits previous dialectometric research, which often found such relationships to be sublinear (Nerbonne & Heeringa 2007).

There is, then, some evidence that overall the lmer modeling process has added value: the relation between geographic and linguistic distances is closer to what one would expect from previous research, almost doubling the explanatory power and yielding a more expected distribution. This, however, should not be interpreted as strong evidence for preferring the clustering in Map 38 over that in Map 36 – even the better of the two fits is still imperfect and quite a bit lower than what is usual in dialectometric research.

5.2.3. GAM-based results

Map 39 displays the results of processing the observed frequencies with negative binomial GAMs followed by noisy hierarchical clustering using Ward's method. The major split is somewhat different here: instead of separating England and Scotland (as in the previous two maps), Southern England, the dark blue cluster, is split off from all other dialects, with the dividing line splitting the Midlands into an eastern and a western part. This group contains the English South without exceptions. However, looking at the corresponding dendrogram (39a), we can see that the groups that exist within this cluster are much less geographically continuous: the next split would separate this large ares into one containing both the very Southeast and -west, and another one containing East Anglia and the more central counties. The Northern English group, in red, spans Wales, the Western Midlands, the North of England and Dumfriesshire in the Southern Scottish Lowlands, but does not include Durham or Northumberland in the Northeast. These two form a group at the English border, together with Selkirkshire, Peeblesshire and Midlothian. The rest of Scotland is divided into the Lowlands (in pink) and the Highlands and Hebrides (in green).

Overall, then, the GAMs result in geographically highly contiguous clusters. This was expected, as the GAM approach assumes that, in the absence of strong evidence to the contrary, dialects that are geographically close are also similar in how they behave. Unsurprisingly, the relation between linguistic and geographic distance is excellent, with a correlation coefficient of $r = 0.61$ and 37.7 percent of the variance explained. Again, a sublinear relation leads to a slightly better result, although the difference here is quite small ($r = 0.62$, 38.2 percent of the variance explained).

5.2.4. Bottom-up analysis: frequency

I now turn to the bottom-up measures as discussed in Section 3.2.3.

Map 40 displays the result of a noisy clustering process using Ward's method on the normalized tag unigram frequencies. Overall, the groupings here are geographically quite discontinuous, but at least partially similar to the previous results. Again, the topmost split is separating Scotland, the blue clusters, from the English counties. The exception is the frequent outlier Middlesex, which groups with Scotland, or more precisely West Lothian. England falls into three groups, only one of which is geographically continuous: the pink cluster, comprising Devon, Somerset and Wiltshire in the Southwest. The remaining two groups show no clear pattern, although the green group contains more dialects from the North of England, namely Westmorland, Durham and Yorkshire.

Let us now turn to bigrams. Map 41 displays the result of a noisy clustering process using Ward's method on the normalized tag bigram frequencies. Once again, we find that the major split lies between the South of England in dark blue, and the more northern dialects. There are again some outliers: Middlesex again clusters with West Lothian in Scotland, and Oxfordshire and Cornwall (in pink) are also far removed from the other Southern dialects. In the north, the remaining two Lothian dialects form the Scottish group, in green, and – with the exception of Lancashire – all counties in the North of England form a single group. In short, there is clearly some heterogeneity in these groupings, but traces of the patterns that were established in the previous sections emerge.

The raw correlations between geographical and linguistic distances are quite similar for unigrams and bigrams: for unigrams, there is relatively little correlation of $r = 0.27$ (7.4 percent of variance explained), while the bigrams show a slightly higher score of $r = 0.32$ (10.2 percent of variance explained).

Overall, the results of frequency-based clustering are unsatisfying in that neither version seems to bear more than a very general resemblance to the dialect areas established either by the previous literature or the manually extracted features as in Sections 5.2.1–5.2.3. Furthermore, while the correlations between linguistic and geographic distances for both uni- and bigrams are higher than the results for the full data set using normalization, they are considerably worse than the corresponding subset of the manual counts (see Section 5.2.1). Two interpretations seem possible. The first possibility is that the aggregate distribution of syntactic features as measured by unigram or bigram frequencies is not distributed spatially. That individual n-grams seem to be, as was established in

Section 4.2, would then be an accidental phenomenon that is not sufficient to influence the aggregate whole. The second possibility is that while bigram usage may be distributed geographically, the frequencies themselves are too noisy to appropriately measure this. Other factors may influence direct frequencies too strongly, and the difference in frequency may not be proportional to the underlying linguistic differences. Crucially, using frequencies directly places heavy emphasis on bigram patterns that are very frequent. For example, `AT.NN1`, an article followed by a singular noun, accounts for about 4.4 percent of the total linguistic difference as measured by bigram frequencies. This pattern actually does exhibit a distinctive distribution, ranking as pattern #21 when ranked according to the number of individual significant differences, and as pattern #150 when ranked according to whole-corpus reliability. At 4.4 percent, however, it seems clearly over-represented in the aggregational result.

5.2.5. Bottom up analysis: reliability

I therefore supplement the frequencies with a measure that weighs patterns according to the reliability of the differences in distribution, and abstracts away both from total usage frequencies and the specifics of the potentially noisy relation of individual frequency differences. Reliability as defined in Section 3.2.3 fits these criteria: First, the per-pattern reliability scores are scaled to the interval between zero and one, limiting the influence that each feature can have on the total aggregational result. Second, instead of the actually observed normalized frequencies, this measure represents how clear the relation between overall usage and subcorpus-specific usage is. Even if a pattern does not have the same frequency in two counties, if that pattern is reliably more frequent in both than in the whole corpus, their reliability scores will usually be similar. Frequency does play an important role, though: not only are the reliability scores ultimately based on the observed frequency patterns, more frequent patterns are more likely to emerge as reliable. This is evidenced by the rank correlation between total frequencies and reliability scores mentioned in Section 4.2. Furthermore, reliability-derived scores were successfully used to identify dialectologically interesting n-grams in Section 4.2, and have therefore proven their usefulness in determining a relevant signal. A final argument for using reliability scores is that in contrast to other methods of determining a more robust signal, such as restricting the analysis to features that are spatially auto-correlated, reliability scores do not take geography into account beyond the grouping of corpus texts into counties. Reliability weighs a pattern that has a multi-county areal distribution the same as one in which the distribution is discontiguous. Thus, reliability escapes some of the circularity inherent in approaches that place geography first and foremost.

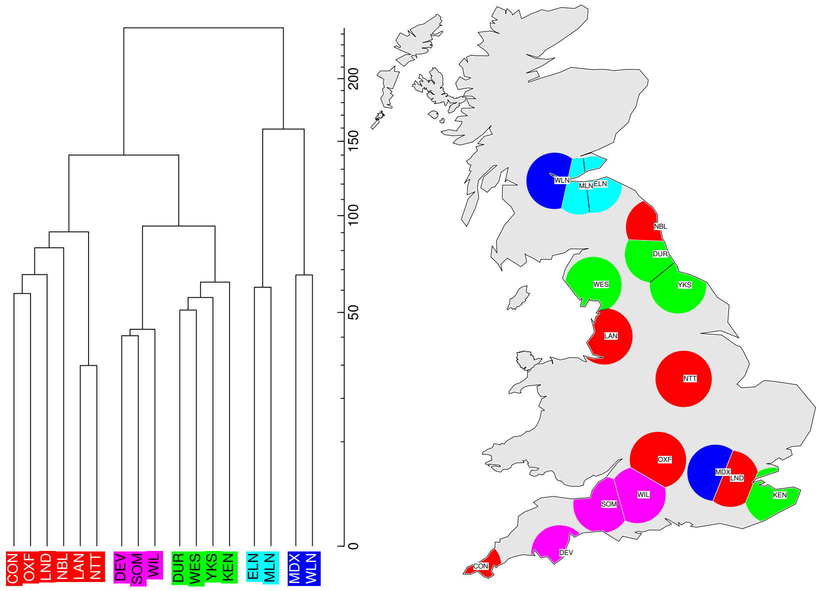
There is one issue with using reliability scores for aggregation. As was noted above, they are scaled to the interval between one and zero, and this is generally a good property as it limits the influence of individual high-frequency patterns. However, it may also lead to very rare features being too influential. For example, a pattern occurring only once will have a reliability of zero for the county in which it appears and a reliability of 0.5 plus some small amount of positive, random noise for all others. While the amount that such patterns contribute to the aggregated distances will overall be small, they are not meaningful and, in large numbers, distract from the pattern inherent in the reliable part of the data. The reliability clustering should therefore be restricted to robust patterns. A simple heuristic is as follows: When establishing pairwise significance values according to the method outlined in Section 3.2.3, we perform 136 different significance tests per bigram[1]. At the customary threshold of $\alpha = 0.05$, we would thus expect $136 * 0.05 = 6.8$ significant differences due to chance. Bigrams that have fewer than 7 significant differences are thus more likely to result from chance. I thus restrict my analysis to those n-grams with at least 7 significant differences. This is the case for 149 of the 221 total unigrams (67 percent) and for 1,899 of the 9,035 total bigrams (21 percent). By restricting the analysis to these tokens, we can eliminate those patterns that are unlikely to be meaningful.

Map 42 displays the result of a noisy clustering process using Ward's method on the significant tag unigram reliability scores. The top split separates the Southern English clusters, in green and pink, from the others; these clusters, however, again are characterized by outliers. First, West Lothian now falls into this group, pairing with Wiltshire. Furthermore, London is not in this group, and instead forms the dark blue cluster together with Nottinghamshire and Lancashire. The remaining Northern English varieties constitute the light blue cluster, and the remaining Scottish varieties the red one.

Map 43 shows the result of a noisy clustering process using Ward's method on the significant tag bigram reliability scores. Here, we find a very contiguous geographic pattern: As usual, the major split in the data is between the North of England and Scotland on the one hand and the rest of England on the other. The north of Britain falls into three groups: The familiar pair of Midlothian and East Lothian as the pink group, the frequent outlier West Lothian paired with Northumberland in dark blue, and Durham, Westmorland and Yorkshire comprising the red cluster. Only Lancashire falls into the periphery of the English South, together with Nottinghamshire in the Midlands. All of the South forms a single group.

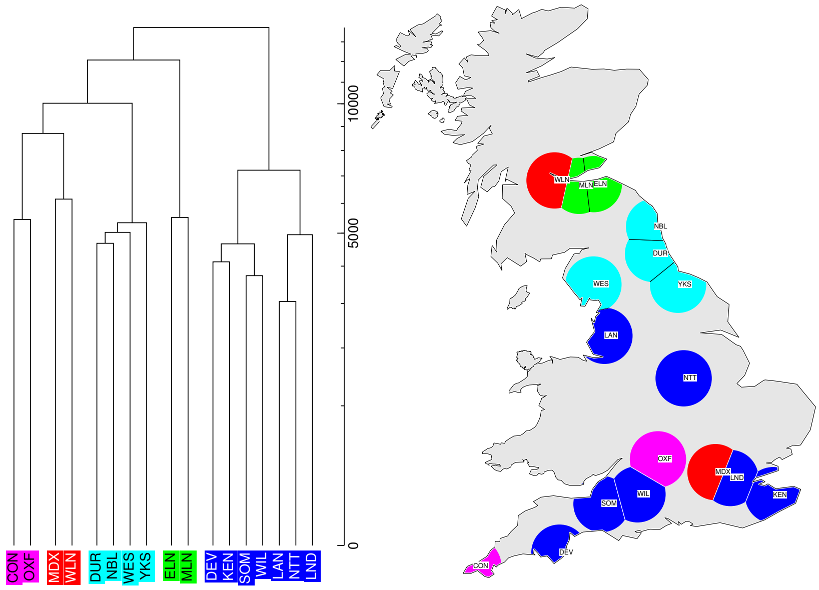
Regarding the correlation between geographic and linguistic distances, we find a con-

[1] There are 17 counties in FRED-S. Comparing each to all other counties would lead to $17*16$ comparisons. This counts each pair twice, so the number of different comparisons is $\frac{17*16}{2} = 136$.



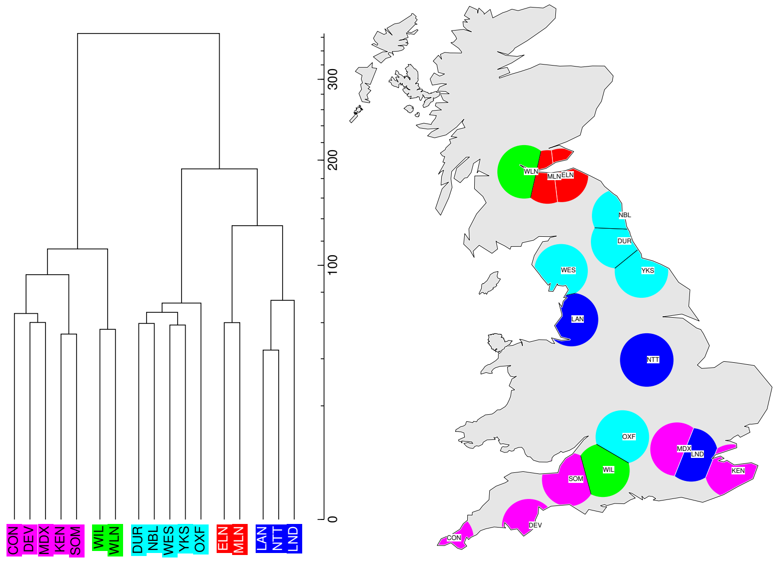
(a) Dendrogram (b) Map

Map 40: Cluster analysis based on unigram frequencies. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.



(a) Dendrogram (b) Map

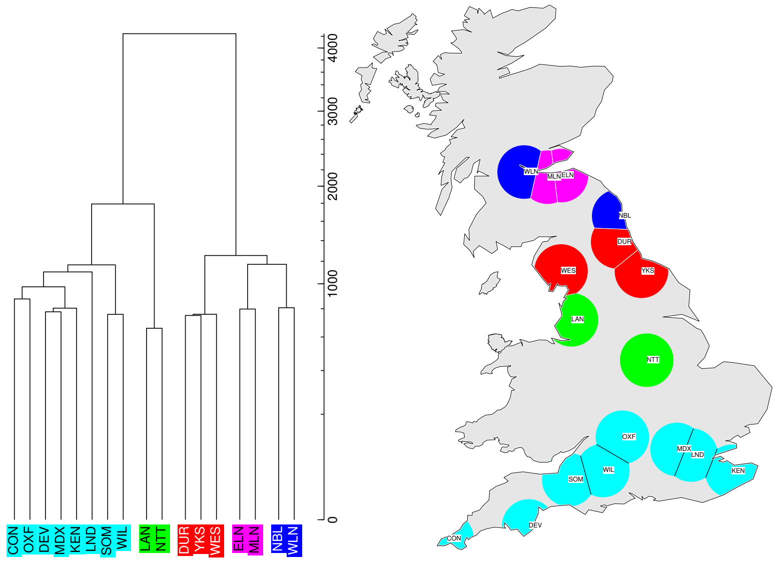
Map 41: Cluster analysis based on bigram frequencies. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.



(a) Dendrogram

(b) Map

Map 42: Cluster analysis based on unigram reliability scores, unigrams with 7 significant differences. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.



(a) Dendrogram

(b) Map

Map 43: Cluster analysis based on bigram reliability scores, bigrams with 7 significant differences. Noisy clustering using Ward's method. Colors indicate group membership in a five-cluster solution.

siderable improvement over frequencies: for unigrams, the correlation is now $r = 0.31$ (9.9 percent of variance explained), and for bigrams, $r = 0.51$ (26.2 percent of variance explained). Comparing these results with those for frequencies, we find that the amount of variance explained is considerably higher for unigrams, and more than twice the previous value for bigrams. The bigram reliability scores lead to correlation values that are comparable to that for manually selected features (cf. Section 5.2.1), although they are still slightly lower.

5.2.6. Interim summary

So far, we have seen that using different analysis strategies and different data sets leads to quite large changes in the dialect areas that emerge from the data. Is everything, therefore, just noise? The answer is no: while the specifics may change, global patterns do exist. The most significant, and the one that appeared in virtually all of the Maps, is that the South of England is different from the British North. Sometimes, individual dialects in the South may group with the northern dialects or vice versa, but the overall trend is overwhelming. Secondly, the northern area divides into Scottish (Lowlands) dialects and those of the English North. This, again, is true for all maps. The precise nature of that split varies, however; interpretation here is complicated by the low number of Scottish varieties in the bottom-up analyses. Nevertheless, in all maps involving the full 31 locations, there is some indication of a transition group between the North of England and Scotland; this transition usually involves Northumberland. Finally, there is little evidence of an east/west split in the South of England: In no map does this split appear as one of the five groups, and neither does it in any of the dendrograms as a subdivision of the Southern English cluster. Variation there is not completely random, however; some pairings appear virtually every time. For example, there is only one map where Somerset does not group immediately with Wiltshire, and that map concerns unigram reliability scores which are by nature based on relatively little – and potentially under-differentiated – data.

This raises the question whether asking for a hierarchical classification is something that the data are unable to handle, because the linguistic reality is simply too complex to fit such a neat picture. If, for example, the transition between the North of England and the South of Scotland is gradual and applies to different features in slightly different ways, we would expect that an analysis that asks to ignore such complications will yield inconsistent results. Minor fluctuations in the importance that certain methods place on individual factors may lead to large differences in the overall pattern. In the next section, I will therefore discuss the results on the basis of analysis techniques that do not enforce a strict hierarchy, namely NeighborNet, or no hierarchy at all, namely continuum maps.

5.3. Intersecting dialect areas and continua

This section provides a closer examination of the distances and groupings discussed in the previous section. Here, splits graphs created using the NeighborNet algorithm as introduced in Section 2.1.2 will be used to allow categorization that is not strictly hierarchical, and thus provides additional gradience in the resulting structures. To allow easier comparison with the cluster dendrograms found in the previous chapter, the same label colors as in the corresponding maps will be used. Furthermore, the networks have been oriented such that, where possible, the northern varieties are placed at the top of the graph.

Figure 5.2 displays the results of using the NeighborNet algorithm on the results of the lmer-based modeling process. With one exception, the clusters that were found in Map 38 emerge as continuous sections in the network; however, as the boxy shapes in the center of the plot show, the hierarchical assumption is not warranted. Nevertheless, subsections of the plot end up comparably compact and tree-like. The members of the red group, all part of the English South, mostly end up very close together, only Middlesex and Suffolk are slightly removed from the core of the group. The pink group containing most of the Scottish Lowlands falls into mainly two parts, with the Kincardineshire-Perthshire pair somewhat removed from the others. Overall, this group also seems relatively compact. The geographically spread out light and dark blue groups show an interesting pattern: In the center of the Network, a relatively large split is apparent, running roughly from Glamorganshire to Somerset. If one were to cut the network there, all dialects from the South of England would be in one half of the network and all Northern and Scottish varieties would be in the other. The only exception to this is Durham, which also grouped with the South of England in Szmrecsanyi (2013). The small green group, consisting only of Northumberland and Midlothian, is placed toward the middle between the English and Lowland Scottish sections of the network. It is also the only cluster that does not form a continuous part of the network. Northumberland turns out to pattern more strongly with the northern Scottish Lowlands than Midlothian does.

In summary, the lmer-based network confirms the separation of English and Scottish dialects as the most significant one. The light blue group acts somewhat like a transition area. The network thus represents both an areal – though, as in the case of the Scottish Highlands and Hebrides grouping with the English North, not necessarily geographically contiguous – signal and a gradient one.

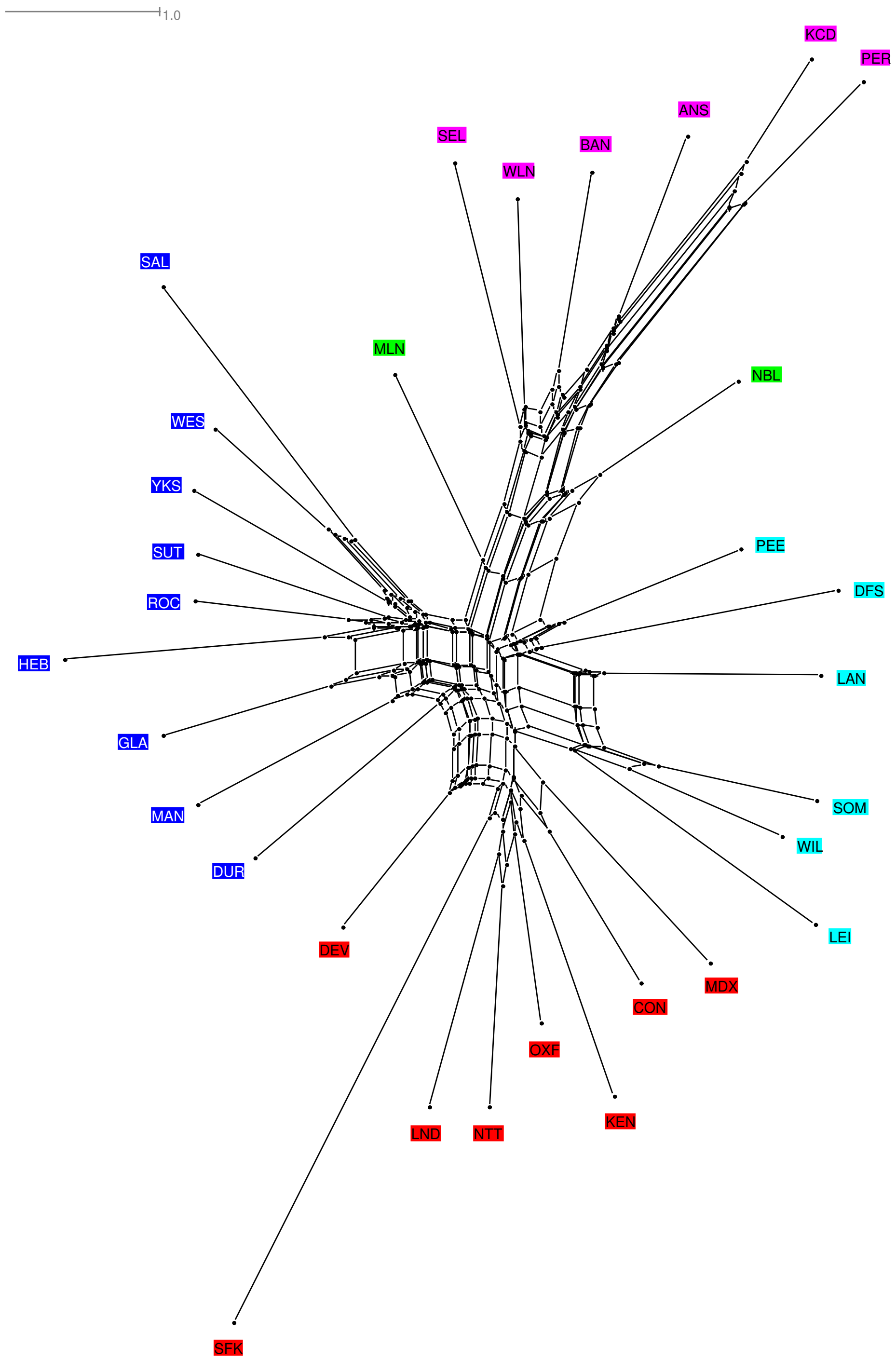


Figure 5.2.: Splits graph based on distances derived from lmer models. Label colors match clusters in Map 38.

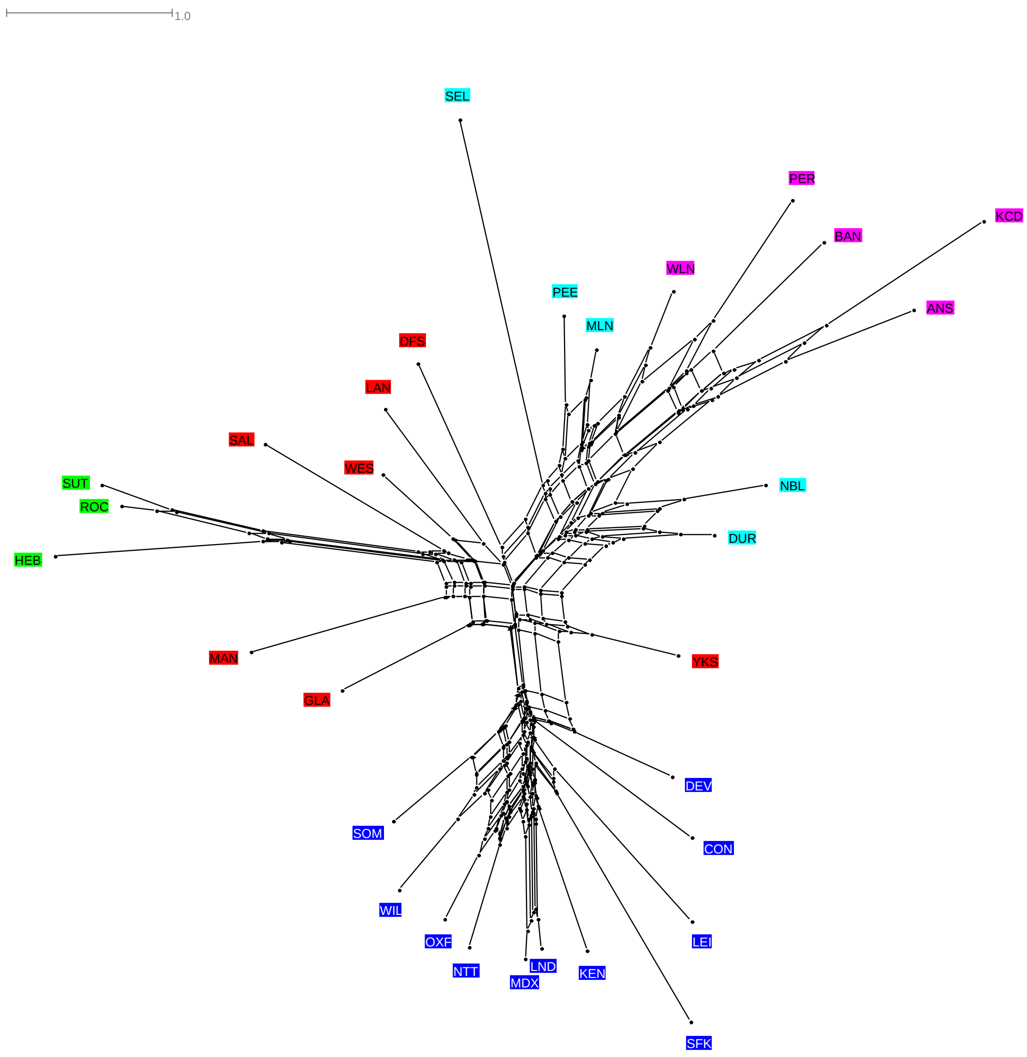


Figure 5.3.: Splits graph based on distances derived from GAM predictions. Label colors match clusters in Map 39.

Figure 5.3 displays the network resulting from using the GAM-derived distances. Again, there is a large split separating the North and the South of Britain; here, the English North forms an intermediate region on the way to Scotland. The Hebrides and the Scottish Lowlands form their own group here, and are removed from the other dialects through a very long split. Still, as in the previous network, this group is placed squarely within the (mostly) Northern English section. Regarding the South of England, we find that the network placement matches the clustering results. Furthermore, from Somerset until Suffolk, the order matches the east/west distribution very well; only Devon and Cornwall in the very Southwest do not follow this pattern. Glamorganshire in Wales and the Isle of Man, which formed a cluster with the western North of England, are quite a bit closer to Southern England than the other members of that group. Yorkshire also has a special place: It is on the other side of the Network. It shows clear relationships to Durham and Northumberland toward its north (in light blue), a fact that the hierarchical clustering could not detect. It is, however, much less similar to the dialects of the northeastern Scottish Lowlands than these two varieties are. The remainder of the light blue group – Selkirkshire, Peeblesshire, and Midlothian – in contrast are rather associated with the west of the English North and the atypical Scottish dialect Dumfriesshire (cf. Szmrecsanyi 2013: Section 7.1.3).

To summarize, the GAM-based network finds that the English South is very different from the remaining varieties, and that the North of England and Scotland form a relatively clear north-south gradient. There, contiguous areal groups can be found, but they form interlocking patterns with other geographically close varieties in ways that are not apparent in the hierarchical clustering.

I now turn to the bigrams, restricting attention to the measure that fared best according to the fit between linguistic and geographic distances, bigram reliability scores. Figure 5.4 displays the resulting network. As with the other networks, the most pronounced split is clearly that between northern and southern dialects. In the South, the original cluster groups are not clearly demarcated by a split in the network; instead three different groups seem to emerge. One of them is the Southwest, where four dialects form a group of their own. London and Kent clearly pattern with the green cluster containing Nottinghamshire in the Midlands and the sole outlier from the North of England, Lancashire. Middlesex and Oxfordshire form the periphery of the southern group. Regarding the north, the English red group emerges clearly, and so does the pink group comprising East and Midlothian. West Lothian, however, is in a clear reticulation: it is similar to the other Scottish dialects in a way that Northumberland is not, and similar to Northumberland in a way that the other Scottish dialects are not.

This can be seen as a gradient transition from English English to Scottish English, as in the GAM network in the previous section. The network perspective thus adds something crucial to the results of hierarchical clustering. Unfortunately, the projection of such interlocking patterns to maps is quite difficult. I therefore turn to *continuum maps*, a technique employed in particular by the Groningen school of dialectometry, to visualize the gradience inherent in the distances. Continuum maps use MDS, a statistical technique for dimension reduction, to boil the distances in feature space down to a number that is easier to handle, namely three (see Section 2.1.2). These three dimensions can then be mapped to the RGB (red, green, blue) color space. Locations that end up close together in the multi-dimensional space therefore have similar colors. RuG/L04 offers three variants of multi-dimensional scaling: *Classical multi-dimensional scaling* (Torgerson 1952), *Kruskal's Non-metric Multidimensional Scaling* (Kruskal & Wish 1978), and *Sammon's Non-Linear Mapping* (Sammon 1969). I follow the recommendation by Heeringa (2004: 160f.) and Szmrecsanyi (2013: 92) to select the method that leads to the best mapping between the MDS result and the original distance matrix. This leads to the selection of Kruskal's method in all three cases. The two model variants have the best match: the MDS based on the lmer values accounts for 93 percent of the original variance, and that for the GAM results for 95 percent. The bigram reliability scores fare worse and only account for 67 percent of the full distance matrix.

Map 44a displays the continuum map for distances based on lmer predictions. Several geographical close areas exhibit quite similar colors: In the South of England, light brown colors dominate. Toward the north, we first find light purple in Leicestershire and Lancashire that grows deeper and more blue. The Scottish Lowlands are mostly in green. Shropshire and Suffolk end up with rather unique colors, testament to their unique position in the lmer predictions due to their good textual coverage.

Unsurprisingly, the GAM predictions lead to the most geographically continuous result, as can be seen in Map 44b. The South of England is uniformly colored in light pink. Nottinghamshire in the Midlands shares this color, and toward the west we find Leicestershire in orange and Shropshire in gray. The North of England is in purple; it starts out relatively reddish in Lancashire and grows more blue toward the East and the North. Selkirkshire, in green, is a clear outlier. The Scottish Lowlands are mostly deep blue, while the Highlands and Hebrides are a very light blue.

The bigram reliability scores in Map 44c, finally, exhibit the least coherence. In general, though, Southern England has relatively deep colors. Lancashire has almost exactly the same green hue as Nottinghamshire in the Midlands. The North of England has light blue-gray colors, and Scotland is again relatively heterogeneous.

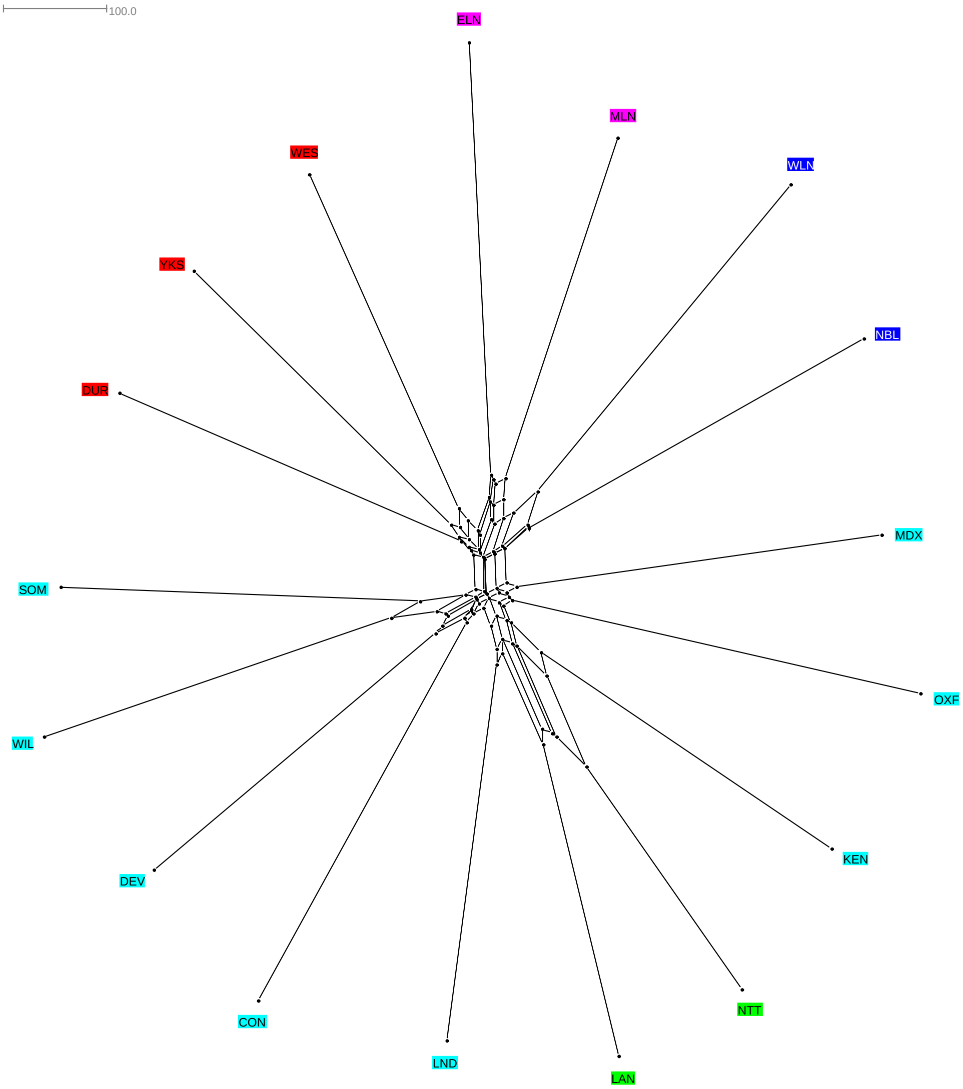
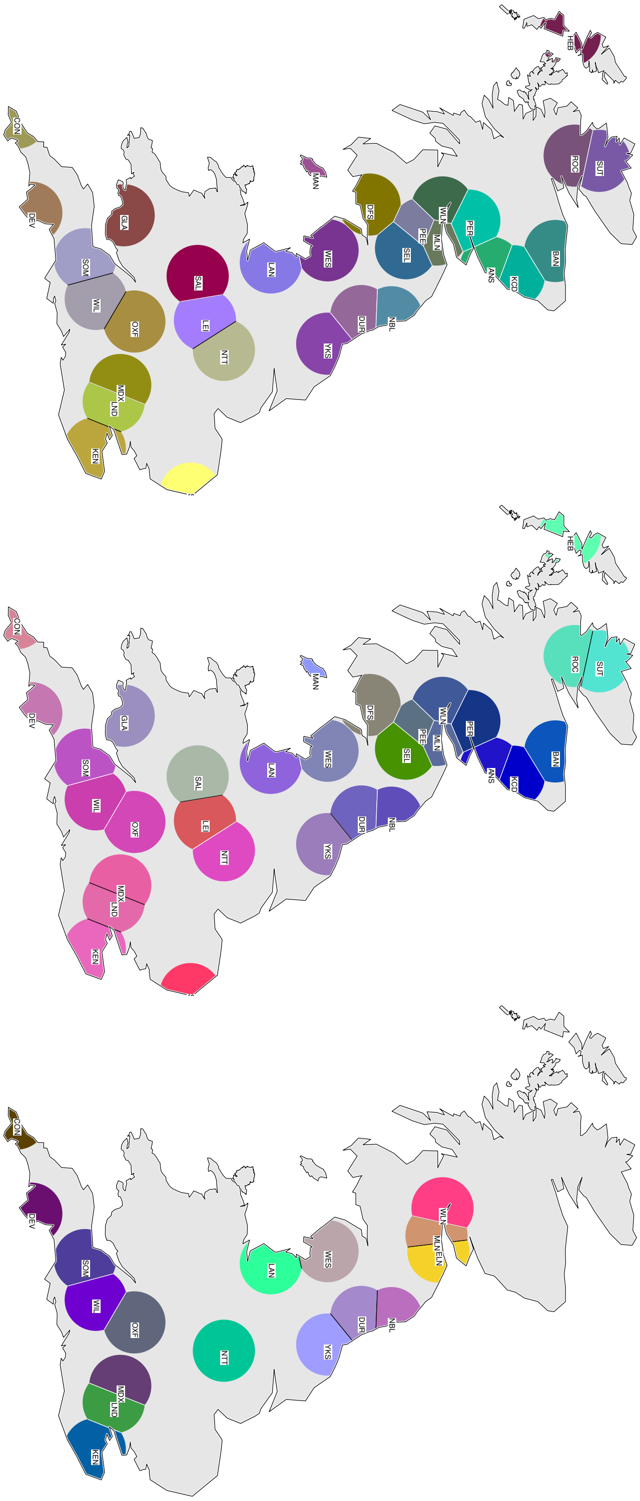


Figure 5.4.: Splits graph based on bigram reliability distances, including only bigrams with at least 7 significant differences. Label colors match clusters in Map 43.



(a) lmer predictions

(b) GAM predictions

(c) bigram reliability scores

Map 44: Continuum maps. Similar colors indicate similar positions in a three-dimensional MDS analysis.

5.4. Chapter summary

This chapter began with a comparison of the effects of the different methods on the individual distances between counties. The results of both model variants were put in relation to the normalized distances and to each other. It was shown that both model types behave as expected: lmer models reduce the linguistic distances for locations with relatively sparse coverage while emphasizing the distances for those with rather large amounts of speakers, while the GAMs tend to level closer points and place greater importance on larger geographic trends.

The distances were then subjected to hierarchical cluster analysis. It was found that the normalization-based measure fares better using only the data where speaker information is available (approximately 90 percent of the total corpus in text size) than on the full corpus, presumably due to the exclusion of some outlier counties where the textual coverage is very low. This improvement can be seen both in the overall fit of geographic distances to linguistic distances and in the geographic spread of the resulting dialect clusters. The lmer-based modeling results lead to an increased fit between geography and linguistic distances, yet the resulting dialect areas were, as in the original study by Szmrecsanyi (2013), often discontinuous. The GAM-based distances had the best fit to geography and all dialect clusters were geographically coherent. For unigrams and bigrams, two measures were explored, one based on normalized frequencies and the other based on reliability scores. It was found that reliability scores lead to an increase in geographic cohesion, both based on correlations and on the distribution of dialect areas.

From a qualitative perspective, several patterns emerged: First, all methods agreed that a fundamental split runs through the data set, separating the northern dialects from the southern ones. The position of the English North shifts somewhat, grouping more with the Scottish varieties in the GAM clusters as well as for unigram reliability scores and for both bigram measures, and more with the English South in the remaining analyses. Some dialects, in particular, tend to shift. Northumberland, for example, is part of the Scottish group in all feature-based analyses except for the FRED-S subset, and part of the English dialect group in all n-gram based analyses except for the one based on bigram reliability scores. Similarly, Lancashire is part of the English North in most feature-based analyses, but tends to group with the South – with Nottinghamshire in the Midlands in particular – in the bottom-up analyses. The Midlands tend to be split, with Nottinghamshire being most similar to the English South and especially the Southeast, while Shropshire tends to fall in with the North of England. Leicestershire, where included, patterns more strongly with Nottinghamshire than with Shropshire. The Southwest of England rarely forms a single cohesive group, although individual pairs, especially the one containing Somerset and Wiltshire, emerge in most analyses.

Next, it was investigated to what degree the tree-like structure inherent in a hierarchical cluster analysis is warranted, and whether the differences between methods are reflected in sub-patterns within the data for a single method. This analysis was performed with splits graphs, using the NeighborNet algorithm. It showed that for lmer model results, GAM results and bigram reliability scores, there was notable non-hierarchical structure in the distances. In the lmer model, a relatively clear North/South split was hidden, reducing the geographic incoherence visible in the cluster map. Furthermore, a gradient pattern was apparent, such that Northumberland is in between the main Scottish Lowlands and English groups. Lancashire is placed on the North side of the British North-South split, but also enters a grouping with some dialects from the Southwest and with Leicestershire in the Midlands. In the GAM-based distances, Northumberland again is positioned in between the English and Scottish group, patterning most strongly with its immediate neighbor Durham. Furthermore, the network uncovers a hidden East-West distribution that holds in most parts of the South of England as well as in the North. The network based on bigram reliability scores found Lancashire to group with Nottinghamshire squarely within the Southern English group. Furthermore, the Scottish Lowlands dialects show a common split that excludes Northumberland.

Finally, the same distances were projected to continuum maps. This largely confirmed the results of the network diagrams. In all maps, Northumberland showed colors that seem intermediate between the Scottish colors and those of Northern England. Similarly, Lancashire exhibited colors that share properties of both the other Northern dialects and those of the South, especially parts of the Midlands.

In summary then, the major groups in Britain are the Scottish Lowlands, the English North, and the South of England. Parts of the North, however, exhibit more characteristics of Scotland or of the Midlands and South than the other dialects there do. The English Southwest is not a clear, cohesive group, although the dialects there are often quite similar in individual pairs.

6. Discussion & Conclusion

This section will begin with a summary of the previous sections. Then, the research questions that were posed in the introduction will be tackled, beginning with those oriented more toward methodology. The structure of dialectal variation in Britain will be discussed next, focusing on the issues that were raised in Section 1.2. I will conclude with a summary discussion of the major themes and suggestions for follow-up research.

6.1. Summary

Chapter 1 introduced the analysis of dialect morphosyntax and the major purpose of this dissertation: to expand and improve the corpus-based, statistical approach to evaluating the geographical distribution of morphosyntactic features, in particular with regard to their frequency. The work presented here stands on the shoulders of pioneering work by Szmrecsanyi (2008; 2013), who introduced a principled methodology for doing aggregate analysis with dialect corpora. Szmrecsanyi's work focuses on British English dialects, and so does this investigation. Therefore, a summary of the existing large-scale classifications of dialect variability in Britain was provided.

Chapter 2 introduced the aggregate approach to linguistic variation in greater detail. Starting with a discussion of how to quantify linguistic material of different data types, namely categorical information, strings, and frequencies, it was then shown how the results of applying methods to several features can be combined into an aggregate measure of the similarities and differences between varieties. A selection of statistical methods that can be used to analyze and display the result were presented. Then, two fields in which the application of such methods has proven fruitful were detailed. The first of these was dialectometry, the statistical investigation of dialect differences. The second was the inference of historical family relations between languages. The chapter concluded with a discussion of three recent approaches for corpus-based analysis of aggregate dialect variation.

Chapter 3 began with a short overview of the major data sources for the present work, namely the Freiburg Corpus of English Dialects (FRED) and its part-of-speech tagged subset FRED-S. The remainder of that chapter discussed the methodology applied here, beginning with the formal explication of Szmrecsanyi's CBDM. Using an example from the data, it was shown that data sparsity poses a problem for this type of analysis. Statistical modeling was proposed as a solution to this problem, and two types of models were presented. The first of these was mixed effects modeling, which leverages the partial pooling effect to reduce the influence of data points with little support. This method was compared with the older normalization-based strategy using simulated data. In this simulation, many parameters were varied, such as the amount of simulated text and the feature frequency. On average, the model performed reliably better, and for rare features the effect was particularly clear. The second type of model was the *generalized additive model* (GAM). This type of model directly incorporates geography by fitting a two-dimensional functional shape over the coordinates of the locations. To do this, GAMs take the surrounding locations into account and try to identify both the overall pattern and the degree to which individual locations diverge from it. The final part of this chapter was concerned which bottom-up analysis, in which the features under study are not pre-selected by the researcher. Instead, part-of-speech combinations that vary by their geographic distribution are allowed to emerge through a permutation-based strategy.

Chapter 4 began with a description of the features, the lmer models and GAMs, and their cartographic representations. Overall, the results tend to harmonize with the existing literature on these features. The choice between negative and auxiliary contraction (Features 34/35), was selected as a case study to investigate how integrating more extensive linguistic annotations into the analysis affects the results of simple modeling. Unsurprisingly, more careful analysis led to an improved result, which was furthermore largely consistent with the existing research on this alternation. Crucially, the results of the simple models matched the more elaborate analysis better than the normalization-based values did. Summary sections then discussed the sociolinguistic and geographic feature distributions. Next, individual part-of-speech patterns that were interesting in their distribution were identified. It was shown that the bottom-up strategy can capture and identify dialectologically relevant features, such as *was/were* variation, *used to* as a marker of habituality, and non-standard *done*.

Chapter 5 presented the results of aggregational analysis on the output of the models and bottom-up measures. First, the effects of model choice were explored. The models behaved as expected: mixed-effects modeling reduced the distances involving locations with little data, GAMs those between close locations. Using hierarchical clustering and their geographic projection, it was shown that normalization and mixed-effects modeling result in relatively discontinuous dialect groups, whereas those for GAMs do not.

Permutation-based bigram reliability scores also exhibited a notable areal structure. Intersecting dialect areas and continua were explored using splits graphs and continuum maps. All in all, British English dialects largely fell into three groups: Scottish English dialects, Northern English dialects, and Southern English dialects. Within these larger groups, smaller sub-groups exist, as do intersections and continua. Depending on the specific analysis individual counties may change their position, but the overall results are similar.

6.2. What do we gain?

The present work consisted of two major parts:

1. a reanalysis of the data used in Szmrecsanyi (2013), using two different strategies:
   - one based on mixed effects modeling using `lme4`, in which geography is represented as a *random effect*, leveraging the partial pooling effect to pull points toward the mean
   - one based on generalized additive modeling, in which geography is represented using two-dimensional smoothers
2. a bottom-up analysis of a part-of-speech tagged corpus

The research questions motivating both parts were as follows:

- To what degree does the amount of available data influence the result of the measurement? Can the influence of this factor be reduced?
- If we can improve the measurement, how does this influence the relationship of linguistic distance to external factors such as geographic distance?
- Do non-geographic factors such as speaker age and gender play a role?
- How do top-down approaches, which start with a list of putatively relevant features and involve a considerable amount of manual selection and coding, compare with bottom-up approaches, which work directly on the data without manual feature selection?
- What do the results of applying these methods on FRED and FRED-S tell us about the structure of morphosyntactic variation in the British Isles?

The remainder of this section will deal with the first four questions, which are primarily about methodology. The dialectologically relevant final question will be discussed in Section 6.3.

6.2.1. On the influence of data availability

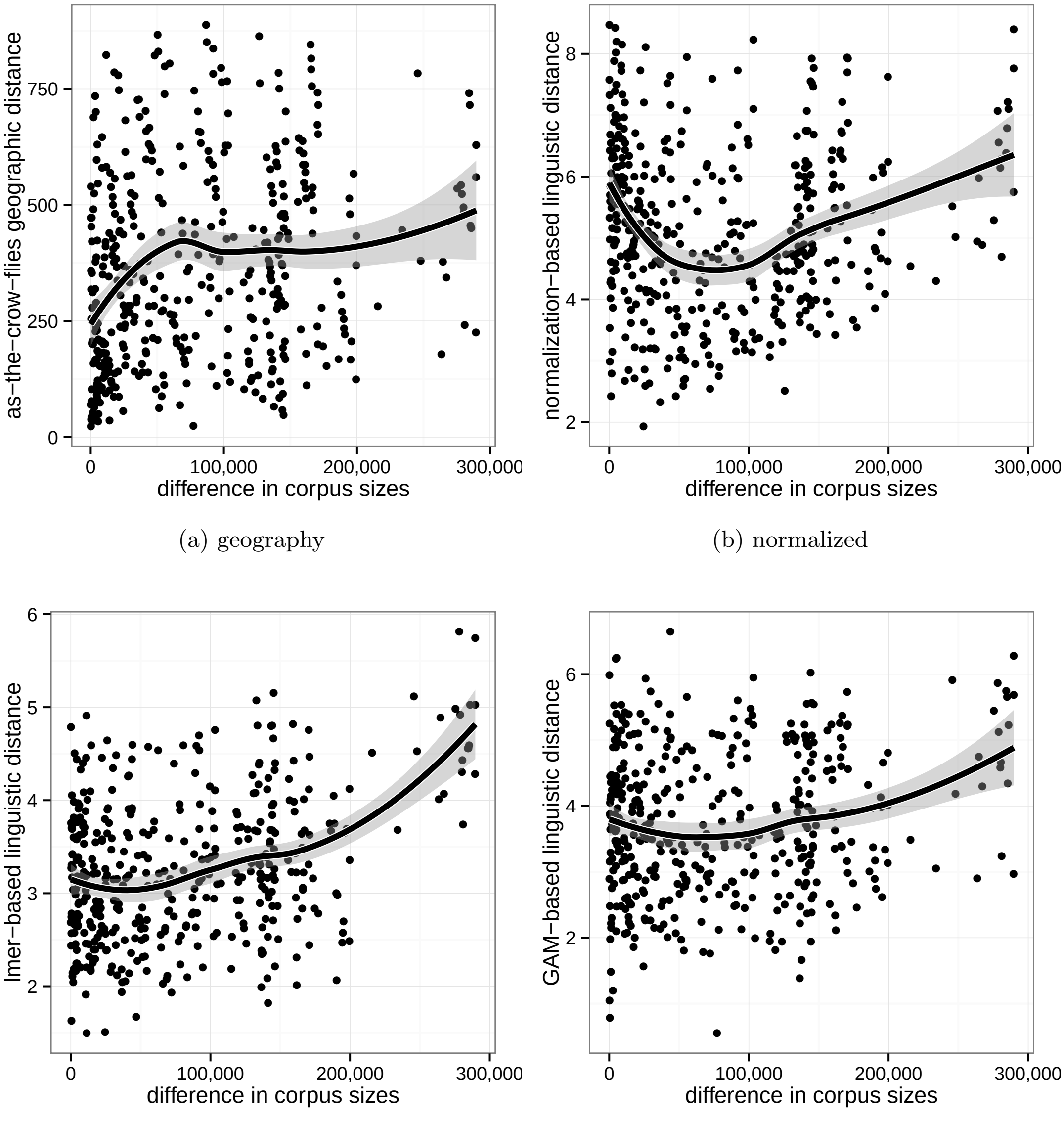
Of the six outlier measuring points in Szmrecsanyi (2013)'s original study, five are among the 10 regions with the least textual coverage in FRED, namely Banffshire, Denbighshire, Dumfriesshire, Leicestershire and Warwickshire. Middlesex alone has relatively good coverage in running text, but relies on only two informants. For the model-based analyses, two of these outliers (Denbighshire and Warwickshire) had to be removed completely due to missing metadata; the other four remain. This allows us to investigate the question whether the amount of data available for a given point has any influence on how large the distance to other points is.

This question is not quite straightforward to investigate, as the distance between two measuring points is one that pertains to that specific pair, but the amount of available data is a characteristic of a single point. We thus need to operationalize data availability in a way that makes sense for pairs of locations. One way to do this is by looking at the differences in subcorpus sizes – if this difference is large, it should be likely that the accuracy of the measurements is different, and thus, if this has an influence on the observed distances, we should see them grow larger as the size difference increases. If there is no effect, we would expect a scatter plot to show consistent scattering of points around a flat trend line. This is the measure that Szmrecsanyi (2013) uses to investigate the problem:

> A potential problem is that normalization carries with it the danger of inflating the effect of freak occurrences due to poor sampling, especially if the corpus is not entirely balanced and textual coverage for some measuring point is comparatively thin. Fortunately, this does not seem, by and large, to be a major problem in the current dataset. [...I]n our dataset large sample size differentials do not generally have an effect on this study's morphosyntax measurements, because sample size differentials do not predict inflated dialectal distances. (26)

Figure 6.1 displays this for the three feature-based analyses.

Let us begin by considering the relationship between subcorpus size differences and geographic distance, displayed in Figure 6.1a, as this is a potential confounding factor. A small trend can be identified: areas with similar amounts of available data also tend



(a) geography

(b) normalized

(c) lmer

(d) GAM

Figure 6.1.: Linguistic distance (y-axis) and corpus size differences (x-axis). Smoother lines indicate overall patterns. Top left plot displays geographic distance against corpus size differences for comparison.

to be geographically close. As the size difference increases, so does geographical distance, up to a size difference of about 100,000 words. At that point the linguistic distance stays level, with a potential uptick for the largest differences. Thus, we would expect that, even if corpus size does not influence linguistic differences, increasing size disparities may still lead to some increase in linguistic distance.

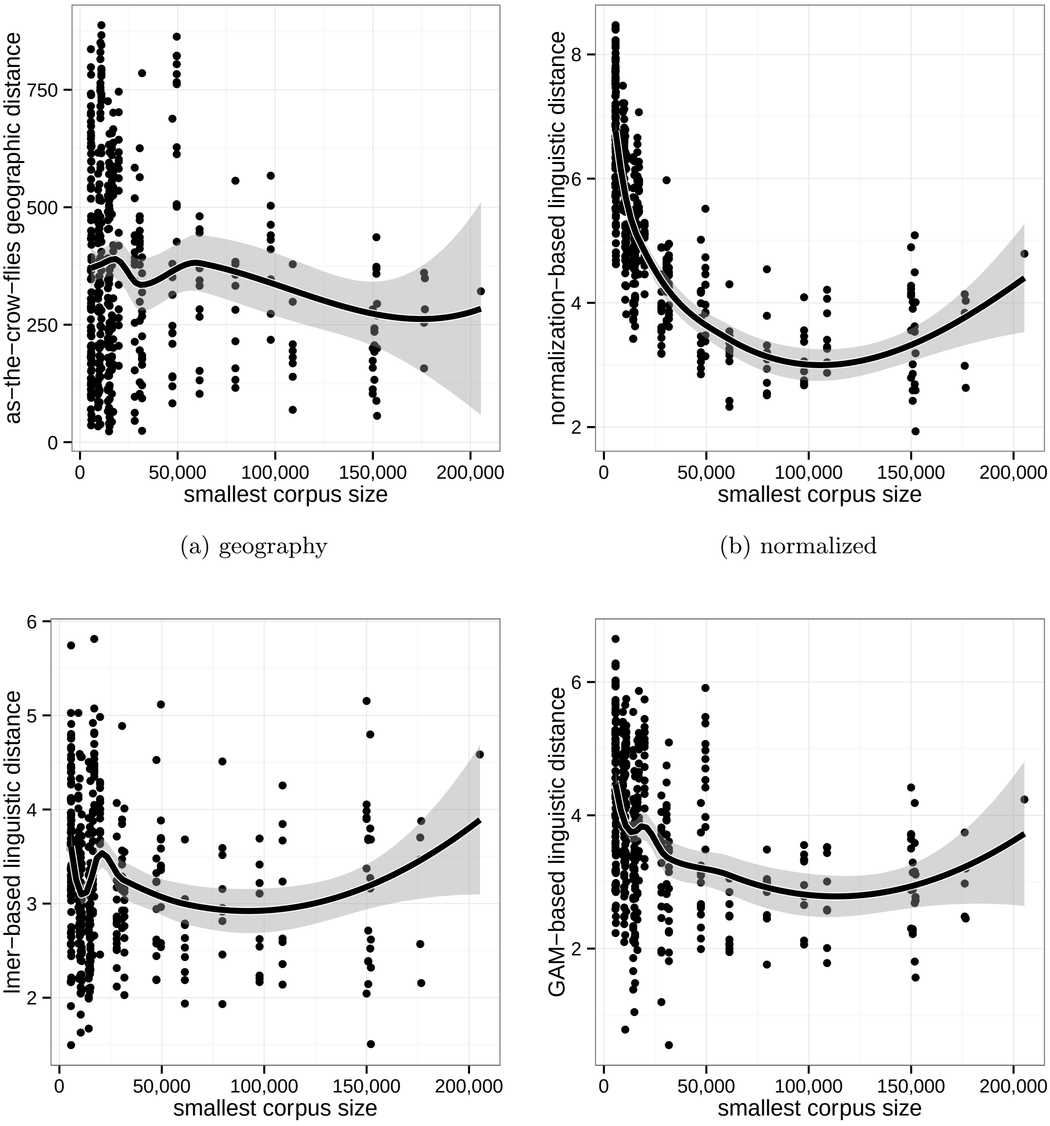
How do the three variants of top-down analysis fare under this measure? The normalized differences in Figure 6.1b show a more or less quadratic pattern, with small differences in size leading to a much larger effect on linguistic distance than slightly larger size disparities. With the size differences of 50,000 words, the distance begins to increase again. Below that threshold, however, the pattern is exactly the reverse of the pattern for geographic distance. This indicates that linguistic distances for subcorpora of similar size may be overestimated. Figures 6.1c and 6.1d show lmer- and GAM-based distances, which exhibit patterns that are similar to the other two plots. Distances based on lmer models behave slightly more like geographical distances, and the GAMs more like the normalized values. Both, however, are considerably flatter for corpus-size differences up to around 200,000 words. It follows that normalization-based distances are affected by corpus-size distances, whereas lmer- and GAM-based measures are less so – perhaps even too little when compared to the geographic distances. The overall effect of corpus size differences, however, was not quite as initially predicted: it is small differences in text size that have an effect on normalization-based distances, not large ones.

This raises the question whether corpus size difference is the appropriate measure to detect the influence of potential “freak occurrences”. Let me illustrate the problem with a small thought experiment. If one were to flip a fair coin ten times, getting three heads would not be a very surprising result, nor would seven heads on a second fair coin. On one hundred flips of two fair coins, however, the same proportions (i.e. 30 and 70 heads) would be much more surprising. With this many trials, the observed counts should be much closer to the expected value, 50 of 100. The (normalized) difference is much more likely to be large for the two coins with fewer flips, as each individual flip contributes more toward the result. This effect does not, however, result from the difference in the number of flips per coin. Within the two sets, the number of flips per coin is the same, and therefore the difference is zero. Comparing coins with unequal numbers of flips, we would expect differences that are smaller than for coins with equal but small numbers of flips: one of the measurements will have lower variance, and is therefore likely to be closer to the expected value. Going back to the actual corpus data, I furthermore note that if both sub-corpora are relatively small, their size difference cannot be particularly large either. Taking this together, we would expect small differences to have large effects and high variability, intermediate size differences to have smaller effects, and the highest size differences to again exhibit large effects, as these necessarily involve the corpora with the smallest total sizes. This is exactly the pattern we see in Figure 6.1b.

Let us thus consider another measure: the size of the smallest corpus. If "freak occurrences due to poor sampling," (Szmrecsanyi 2013: 26) are a problem, then pairings involving at least one small corpus should have very large distances, and as minimal corpus size increases the distances should become smaller. The results are displayed in Figure 6.2.

Again, we begin by considering geographical distance (Figure 6.2a) as a control. Here, we see that when both corpora are rather large, they also tend to be relatively close geographically, as was shown in Map 2a (page 45). Overall, however, there is almost no correlation ($r^2 = 0.02$). The picture is radically different for normalized distance, as Figure 6.2b shows. When the smallest corpus is particularly small, the linguistic distances are very large and very variable; as the minimum size increases, the linguistic distance decreases. For the highest minimum sizes the distance increases again, but that section of the plot is based on too few points to properly evaluate this. The correlation is very high, with r^2 being 0.61. In other words, over 60 percent of the variability in the normalized distances can be explained just by knowing the number of words in the smallest corpus involved. The lmer differences, in contrast, are a little wiggly, but exhibit little correlation ($r^2 = 0.02$). Instead, we find a slight trend on a similar measure: distance increases as maximum text size increases ($r^2 = 0.08$). The GAM values, finally, are slightly more similar to the normalized values in that they decrease as corpus size increases, but the effect is much weaker and the r^2 value here is only 0.16. The maximum number of words does not exhibit notable correlation ($r^2 = 0$).

In summary then, Szmrecsanyi (2013)'s claim that "freak occurrences [are] not a major problem" (24) seems to be too optimistic: using normalization, the amount of data available for each location does have a notable effect on the linguistic distances it is involved in. This is not the case for distances resulting from lmer predictions, and only to a much lower degree for those resulting from GAM predictions. We can thus consider the model-based approaches successful in their goal of reducing the influence of one confounding factor, although neither removes it completely. The next section will discuss whether this finding has an effect on how geographic and linguistic distances relate to one another.



(a) geography

(b) normalized

(c) lmer

(d) GAM

Figure 6.2.: Linguistic distance (y-axis) and minimal corpus size (x-axis). Smoother lines indicate overall patterns. Top left plot displays geographic distance against minimal corpus size for comparison.

6.2.2. On geographic and linguistic distances

The exact nature of the relationship between geography and linguistic distance has received considerable amount of attention from dialectologists and especially dialectometrists. Nerbonne & Kleiweg (2007) formulate the central idea as follows:

> **Fundamental Dialectological Postulate**: Geographically proximate varieties tend to be more similar than distant ones.

Of course, this is a statistical generalization; following Saussure (1916 [1983]: 271), "[g]eography alone can have no influence upon a language." It is only as a proxy for other causal factors that geography appears as a determinant. Yet, the evidence that this generalization holds is strong, not only in geolinguistics, but across the geographical sciences. Tobler (1970), to wide acclaim, formulated this as the first law of geography: "everything is related to everything else, but near things are more related than distant things".

Regarding dialectal data alone, the evidence from around the world is overwhelming. Nerbonne (2013) provides an overview of several studies using pronunciation data from Bulgaria, Germany, the Netherlands, the United States and Gabon Bantu. Across these data sets, the influence of geography as a predictor of linguistic distance ranges from 16 to 37 percent. More related to the topic at hand, Spruit et al. (2009), in a study of Dutch dialects, found that high correlations between linguistic and geographic distances is not restricted to phonetic data: distances based on pronunciation can be attributed to 47 percent to geography (as measured by the r^2 value), for syntactic differences to 45 percent, and for lexical variation still to 33 percent. Moreover, the general pattern applies to British English dialects as well: Shackleton (2010: 167) reports that "roughly half" of the variation in British English dialect phonetics can be attributed to geography. This can be increased even further by allowing the precise effect on individual dialect regions to vary.

Another consideration pertains to the shape of the relation between geography and language. Séguy (1971) first reported a sub-linear relationship, i.e. as geographic distance increases, the rate at which the linguistic distance increases becomes smaller. Later research, especially by the Groningen School, has found relationships of this nature in many different data sets (e.g. Nerbonne & Heeringa 2007, Nerbonne 2009). Another well-known model is Trudgill's *Linguistic Gravity Hypothesis* (1974). It seeks to explain the linguistic relationship between two sites as a function of their distance and of their population sizes, in analogy to Newton's gravity. More specifically, the influence of site i

on site j is:

$$I_{ij} = S_{ij}\frac{P_i P_j}{d_{ij}^2} * \frac{P_i}{P_i + P_j}$$

where S_{ij} is their preexisting similarity, P_i and P_j are the respective population sizes of i and j, and d_{ij} is their distance. This hypothesis is based on the observation that larger sites have greater influence, and that innovations first spread across influential sites akin to “skipping a stone across a pond” (Chambers & Trudgill 1998: 166). The model has had considerable success in explaining the distribution of individual features, but its adequacy in the aggregational perspective has largely failed to surface: Nerbonne & Heeringa (2007) note that their data fails to show the quadratic effect of distance that the Linguistic Gravity Hypothesis posits.

In contrast, Szmrecsanyi (2012; 2013) reports very different findings for distances based on corpus frequencies of morphosyntactic features. First, geography explains very little of the aggregate variation. Three different operationalizations of geographic distance were tested, including the straightforward as-the-crow-flies distance between sites and estimates of both walking time and modern travel time (i.e. using motorcars and modern infrastructure). These estimates were derived using the route finder of the British version of Google Maps (for details, see 2013: 103f.). Neither measure fares particularly well: as-the-crow-flies geographic distance can only explain 4.4 percent of the total variance in linguistic distances, and the travel time estimates only lead to a small improvement, explaining 6.8 percent for walking time and 7.6 percent for driving time. Szmrecsanyi then restricts his attention to two subsets of the distances, comparing only English dialects in the first and only the Scottish Lowlands in the second. For all operationalizations of distance, there is no significant correlation within England. For Scotland, however, the geographic signal is much more informative, explaining 33 percent of the variance using as-the-crow-flies distances and a full 39.4 percent using travel time by car. The usually strong influence of geography thus only held for the Scottish Lowlands, and only to a very low degree for the total data set. The shape of the relation is, again with the exception of Scotland, also matched better by a simple linear pattern than the expected logarithmic one (Szmrecsanyi 2012).

Then, Szmrecsanyi (2013: 105f.) explores the Linguistic Gravity Hypothesis, drawing on English, Welsh and Scottish census figures from 1901 and the travel time estimates as above. He finds that linguistic gravity emerges as the best predictor for both the total data set and the Scottish Lowlands, explaining 24.1 and 46.5 percent of the variance. This relationship cannot be reduced to the component parts of the gravity values, as both squared travel time and the population product by themselves can only account for less than ten percent of the variance. Again, this is in stark contrast to other results in dialectometry.

Szmrecsanyi (2013: 159f.) considers three possible explanations for this pattern. One is that the dialects of Britain could be different from those in other areas in that they are less structured in a geographic manner. This claim can be rejected based on the findings from Shackleton (2010) – at least based on pronunciation, Britain seems no different from other regions around the world. Could it be that morphosyntax is distributed less geographically? Szmrecsanyi denies this as well, based both on the results of Spruit et al. (2009) and on the fact that theoretical arguments about the lower diffusability of grammar have turned out to be at least partially unfounded. His final explanation is a methodological one: the distances that derive from dialect atlases essentially rely on "data reduction" (cf. Wälchli 2009) in that they abstract away the variation that may exist at individual sites. This argument is corroborated by an experiment in which a reduction technique is applied to the data set, replacing the absolute frequency values with either frequency rankings or frequency categories, and finds that the influence of as-the-crow-flies distance increases to about 7 for the first and 9 percent for the second. Similarly, selecting only features that exhibit a geographic distribution by themselves yields an effect of geography accounting for 14.6 percent of the variance.

As we have seen in the previous section, however, there is another explanation: the linguistic distances derived from the corpus scores could be too heavily influenced by factors external to language or geography, namely the low amount of corpus material at several of the sites. Let us test whether this can account for the results observed by Szmrecsanyi. We will walk through three major results in turn: the effect of linguistic gravity, the differences between England and Scotland, and finally the strength and the shape of the relationship between geography and aggregate linguistic variation.

We established in the previous section that there is no correlation between minimum text size and geographic distance. Is the same true for linguistic gravity? As it turns out, it is not: correlating both, we see that the lowest number of words in a pair predicts their gravity score to about 22.3 percent. Moreover, there is a notable difference between England and the Scottish Lowlands with regard to this. Considering the Scottish Lowlands alone, we find a much stronger correlation, accounting for 37.3 percent, while there is no correlation for England at all ($r^2 = 0$). This very closely mirrors the relationship between the linguistic distances using normalization. For the modified data set used here they turn out slightly different from Szmrecsanyi's original results, with an increased fit for all locations of 30 percent and a small reduction for the Scottish Lowlands to about 31.9 percent. In other words, not only is the pattern of correlation to morphosyntactic distance for linguistic gravity similar to that for a crude measure of data availability (60.6 percent total and 43.6 percent in Scotland), in absolute numbers it even fares worse.

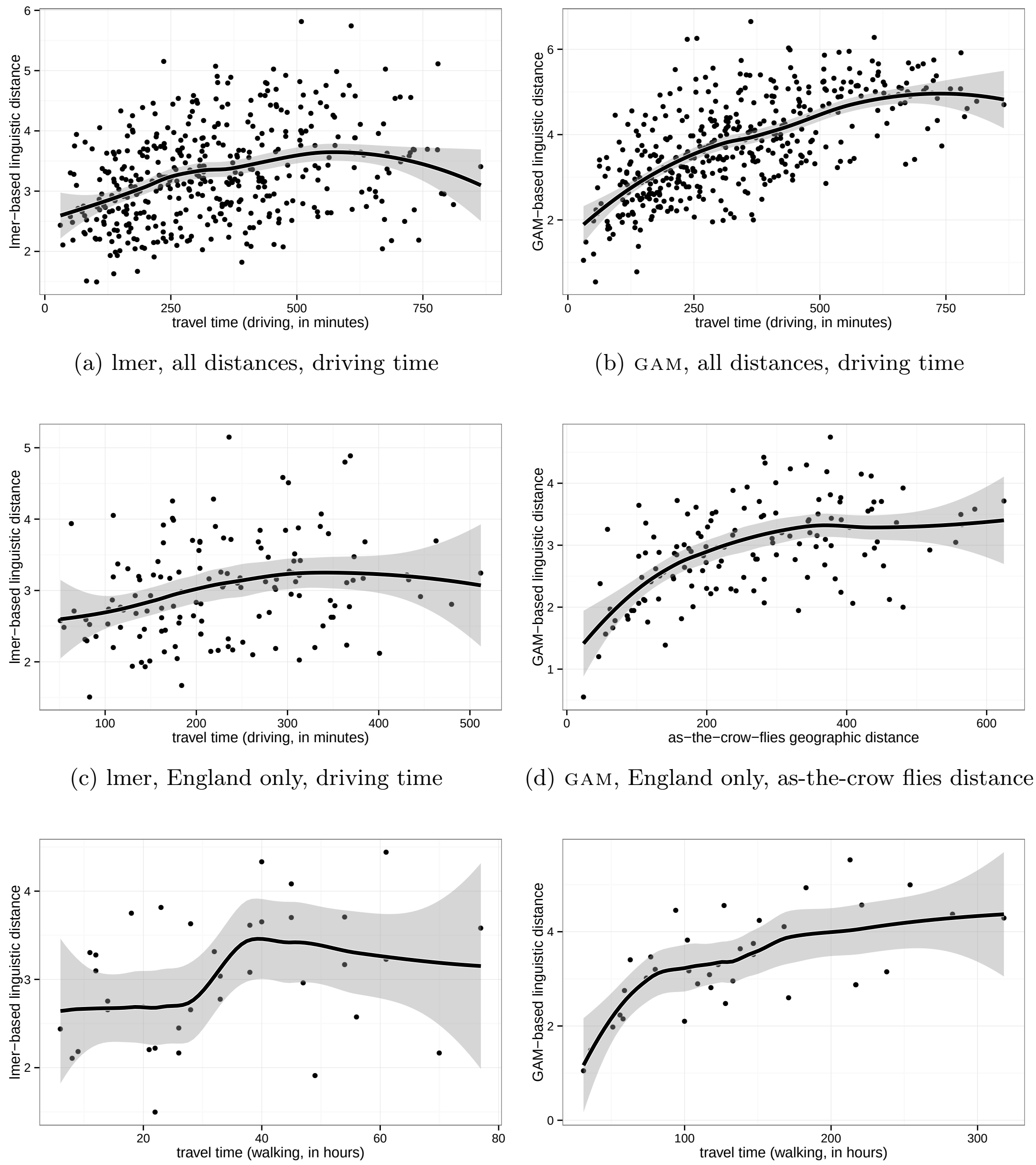
This relation can also explain the difference between England and Scotland not only with regard to the influence of linguistic gravity, but also to that of geography. As we have seen, overall there is no correlation between minimum text size and geography, and the same is true when considering only England. For the Scottish Lowlands, however, there is again a strong signal, accounting for 34.4 percent of the variance. Considering the strong effect of that factor on the linguistic distances resulting from normalization, it is thus not surprising that Scotland fares so much better than England does: the distribution of texts in Scotland simply resembles both geographic distance and linguistic gravity much better.

The distances resulting from lmer modeling do not exhibit such a correlation to minimum data availability, but a minor one to maximum data availability instead. Those resulting from GAMs exhibit the relation to minimum text size only to a much smaller degree. Does geographic distance fare better under these circumstances? As it turns out, it does. Concerning all distances resulting from lmer models, as-the-crow-flies distance can explain 10.1 percent of the variance, a notable increase over the 4.9 percent of the unmodeled variants. England again fares worse with respect to the influence of geography, but shows a similar increase, from zero to 4.6 percent. For Scotland, the strong correlation that was observed by Szmrecsanyi (2013) is reduced to 8 percent, which is consistent with the hypothesis that the original finding was an artifact of data availability. Let me now address the shape of the relationship. With the exception of Scotland, these values can be increased further by considering the logarithm of the distance, although the difference is numerically rather small. For the full data set, a sublinear curve explains 12 percent for the total data set and 6.5 percent for England. The travel time measures do not change these results by much. Both in the total data set and in England only, the logarithm of driving time improves the values by about one percent. In Scotland it is the logarithm of walking time that fares best and increases the fit to 12.4 percent. In short, after the lmer models remove the strong influence of corpus size, we achieve a stronger effect of geography, and one that is consistent with the sub-linear relationship reported in the literature. In absolute numbers, the effect is, however, still much smaller than those reported in atlas-based studies.

I now turn to the GAM-derived distances. On the total data set, as-the-crow-flies distance explains 37.7 percent of the variance, a dramatic increase, and a value that seems more in line with the expectations from other dialectometric research. For Scotland alone, this value remains essentially unchanged, while England shows a less strong yet still respectable value of 24.8 percent. Once more, however, we can improve on this by assuming a sublinear relationship and by using travel time measures. The most explanatory metric for the whole data set turns out to be the logarithm of driving time, which can account for 44.3 percent of the variance. Similarly, the logarithm of walking time works best for Scotland at an explanatory power of 54.2 percent. Only the distances within England best match the (sublinear) as-the-crow-flies distances at 33.4 percent.

Figure 6.3 visualizes these distributions. The black lines are LOESS smoothers indicating the trend in the data. The GAM-derived distances (Figures 6.3b, 6.3d and 6.3f) show the clearest patterns, and all of them are sublinear in nature. The patterns for the lmer-derived distances are less obvious, but still show a sublinear curvature for all distances and for England. Only the values for the Scottish Lowlands have a notably different pattern: here, for small distances up to about 30 hours of walking time there is no relation between linguistic and geographic distances. For travel times longer than 30 hours, the average linguistic distance increases quickly, then levels off at about 40 hours travel time. This may be an artifact of data availability – the plots containing only the Scottish Lowlands are based on the smallest number of distances.

Where does this leave us with respect to the influence of geography? Is the claim by Szmrecsanyi that "[c]ompared to corpus-based and frequency-centered approaches, atlas-based approaches overestimate the importance of geography" (2013: 160) false? Not necessarily. While the distances derived from lmer-based and GAM-based predictions do fit the pattern suggested by the literature much better than Szmrecsanyi's, in doing so they effectively use data reduction. The effectiveness of that method in boosting the effect of geography was already confirmed. The precise manner of, and motivation for, data reduction is different in this study, but that does not change the fact that the models are, by nature and design, less sensitive to some types of frequency differences, especially in low-data situations. The question, then, is whether doing so is warranted; and that is a question that cannot be answered easily. From the linguistic perspective, the maps that result from the GAM process seem meaningful, in that they largely match what is reported in the literature pertaining to the modeled features. Furthermore, the results of the case study presented in Section 4.1.2 suggest that the values resulting from simple models may reflect those that a linguistically more sophisticated analysis produces much more closely than the normalized values do. If one accepts the Fundamental Dialectological Postulate, it makes sense to require that good evidence be brought to the table before believing that two proximate varieties are not similar. Then, however, the results cannot be used to argue for that postulate, as the conclusion is already assumed. The lmer-based method, which makes weaker assumptions – only "everything is related to everything else" and not



(a) lmer, all distances, driving time

(b) GAM, all distances, driving time

(c) lmer, England only, driving time

(d) GAM, England only, as-the-crow flies distance

(e) lmer, Scottish Lowlands only, walking time

(f) GAM, Scottish Lowlands only, walking time

Figure 6.3.: Linguistic distance (y-axis)and geographic distance (x-axis). Smoother lines indicate overall patterns. The geographic distance measure chosen is that with the overall best correlation.

"near things are more related than distant things" in Tobler (1970)'s parlance – finds a geographic signal that lies in between these methods but closer to the unmodeled variant. Yet, the restricted assumptions may well not be enough – why should data from Angus not give us more information on what Banffshire is like than data from Cornwall does?

Nevertheless, I argue that the influence of corpus size on linguistic distance is a real problem for the CBDM enterprise, and that, at some level, the researcher will need to pay the "price" (Szmrecsanyi 2013: 165) of abstracting away from variation in her data. This abstraction can happen by various means, whether simply by ranking as in Szmrecsanyi's experiment, by means of Getis-Ord *Gi*∗ hotpot analysis as performed by Grieve (2009), or via probabilistic modeling and permutation-based reliability as presented here. If the analyst is unwilling to do so, she runs the risk of confusing the forest with the branches of a tree.

6.2.3. On non-geographic factors

Dialectometrists have recently begun to include social variation explicitly (e.g. Wieling et al. 2011, Wieling 2012, Hansen 2011). As Nerbonne & Heeringa (2007) remark:

> Finally, and especially given all of the attention which has been paid to social factors in language change [. . .] it would be most attractive to analyze data which has been collected to systematically catalogue variation over a range of extralinguistic variables, including at least geography, class and sex. This would allow a more direct comparison between the roles of geography and other social factors. (292, references omitted)

FRED is, by design, not the corpus required for this investigation. Nevertheless, the analyses in Chapter 4 have shown that some of the variation in the data can be captured just by knowing age and gender of the speakers. Furthermore, this variation is patterned, confirming the wide consensus in sociolinguistics and dialectology that, in general, female speakers use more standard variants, and older speakers more non-standard variants. In some cases, the estimated differences were quite large.

Would these differences strongly affect the linguistic pattern that emerges, though? Let us turn to the models to attempt an answer to that question. When calculating the GAM predictions, we specified that for each location the predicted data should come from a male speaker that had the average speaker age. We can rerun the same process, simulating for variants in speaker age or gender. I repeated this process 100 times each for the following variants:

- setting (centered) speaker age to a random number drawn from a normal distribution with the standard deviation equal to that in FRED (per county)
- randomly choosing some speakers to be female, with a probability equal to the proportion of female speakers in FRED
- the above combined

We can then investigate the correlation between distances derived from default speakers (i.e. male and average age) and what a more varied corpus would look like, according to the data. If the correlations are high, this is an indication that – at least for the data set and methodology used in the present work – sociolinguistic factors are not a central influence on the variability in the data, and could in principle have been removed from the analysis. If the correlation is low, sociolinguistic factors matter. As it turns out, the former is mostly true. For age, knowledge of the default speaker explains, on average, 98.8 percent of the variance in the randomized data. The effect of gender is more notable, but at 95.0 percent still rather negligible. Their combination does not change much either, as r^2 still sits at a comfortable 0.93.

Can we also find this pattern in the normalized linguistic distances? Regressing linguistic distance on the differences in mean age and gender proportion as well as their interaction leads to a significant, yet very weak signal. Keeping as-the-crow-flies distance as a control, we find that both larger age differences and larger gender proportion differences increase the average distance while the combination of both reduces it. These predictors account for about 2.4 percent of the variance. In other words, variation along social axes has an effect in this data set, but from an aggregated perspective it does so only to a negligible degree.

6.2.4. On bottom-up versus top-down analysis

Let us now turn to the results of the bottom-up n-gram analyses. How do they fare, compared to the normalized measure? First, I note that the original CBDM methodology fares much better when only considering FRED-S, at least as far as the relation between linguistic and geographic distance is concerned: knowledge of one explains a full 27.6 percent of the variance in the other. Given that many of the counties with particularly thin coverage are not included, this is not surprising: as we have seen, CBDM works best when there is enough data. This sets a high bar for the fully automated analyses. And it is one they fail to pass, particularly as far as n-gram frequencies are concerned.

Perhaps unsurprisingly, unigram frequencies fare relatively badly, with a correlation to geography of 7.4 percent. Bigrams do slightly better at 10.2 percent. These results are consistent with Sanders (2010), who in a survey of different parameter settings found correlations of about $r = 0.1$ to $r = 0.3$ for unigrams and trigrams, i.e. r^2 values ranging from about zero to 10 percent of explained variance. Distances based on reliability show a more encouraging sign of 9.9 percent for unigrams and 26.2 for bigrams. In other words, this again confirms Szmrecsanyi (2013)'s observation that frequency-based measures often show little effect of geography, but that abstractions based on frequency can fare a lot better.

How are these distances affected by data availability? First, it should be noted that, as with the correlation between geographic and linguistic distances, the normalization-based metric fares better on FRED-S. Nevertheless, with an r^2 value of 0.485, there is still a very strong relationship between the two. For bigram frequencies, the problem is worse, with minimal corpus size explaining 67.2 percent of the variance in the resulting distances. Bigram reliability score distances again are less affected: their r^2 value is 0.16, virtually identical to the correlation of GAM-based distances and minimum corpus size on the complete data set. In short, and at least on this data set, the permutation-based approach can ameliorate imbalances in the data to a degree.

Pure bottom-up approaches alone seem insufficient for dialectometric purposes. What then is their advantage? Of the CBDM approach, Szmrecsanyi (2013: 165) states that

> [i]t can, in principle, just as well be applied to variability in modern dialects and accents. And this is a desideratum that is high on the agenda. David Britain has noted that "there are huge gaps in our knowledge of the present-day grammars of varieties in England" (Britain 2010: 53), and we believe that CBDM is a methodology that advertises itself for addressing these gaps from the bird's-eye perspective, in tandem with more traditional variationist analysis methods designed to cover the jeweler's-eye perspective.

If that is the case for a very labor-intensive approach, it should also be true for a linguistically much less sophisticated, but faster one. Especially in cases where there are huge gaps, quick methods that yield both a first overview and an automated selection of potentially interesting features for further investigation should prove useful. The examples in Section 4.2 have shown that, using the techniques proposed here, bottom-up approaches can achieve that. Furthermore, even if the overall fit to geography is less than one would achieve using atlas-based or top-down corpus-based measures, the dialect groupings that emerge from it seem to be meaningful. This has been shown in Chapter 5, and will be picked up once again in the next section.

However, the greatest allure of such methods is that they scale well to larger amounts of data. The assignment of part-of-speech tags is a process that can be done automatically with a rather high degree of accuracy, and nothing stands in the way of fine-tuning the involved algorithms to the specific characteristics of dialectal data. Similarly, determining significant differences and distances based on frequency and reliability can proceed without further input. If, in the future, appropriate data becomes available on a large scale, such fully automated techniques may well become very useful, supplementing the bird's eye perspective and the jeweler's eye perspective with satellite imagery.

Let me conclude this comparison of top-down and bottom-up approaches by raising the question why I discuss both together in the same work. After all, the two seem quite different in both the precise nature of the data used and in their methodology. Put another way, both methods are clearly corpus-based and aggregational, but what makes the bottom-up approach probabilistic in the way that the top-down approach is? And, I must admit, this question is not easy to dismiss. The two are certainly not so intrinsically linked that the connection is automatic or necessary. Nevertheless, I would like to argue that not only do the two go well together in that they reinforce and complement each other, but they are also related in meaningful ways, even if this is not immediately obvious. The central property of the top-down methods introduced in this work is that they abstract away from pure observed frequencies into values that remain thoroughly driven by frequencies, but are more robust with regard to the noise inherent in corpus data. Similarly, for the bottom-up approach, we find that reliability scores, which are derived from frequencies, yield better results than frequencies alone, and are less affected by data imbalances. While there are large differences in how the models and the permutation-based approach achieve this, there is also an underlying similarity. The models calculate the probabilities of a certain grammatical feature, or its occurrence rate, which is in essence the probability for each word to be an instance of this feature. The permutation-based approach yields the reliability measure, which is directly influenced by the probability that a random corpus has a higher feature probability. Whereas the previous applications of permutation-based techniques only permuted to evaluate frequency differences in terms of significance, I replace the frequencies with, essentially, probabilities. Furthermore, I show that doing so not only increases the signal, but also reduces the harmful effects of data imbalances. In other words, the reliability scores perform the same job as the models do, just in radically different ways. They are not probabilistic in the sense that the top-down methods are, and do not yield anything that can be interpreted as a probabilistic model, but I think the term probabilistic is still appropriate for them. Probabilities as a way of dealing with uncertainty in the face of data sparsity is the central idea that unites the two approaches.

6.3. On the dialect landscape of Britain

In the previous section, we have seen that the methods proposed here can reduce external influences compared to a strategy based on normalization; furthermore, this necessitates changes to the interpretation of some geolinguistically relevant questions. What, then, do these results tell us about morphosyntactic variation in England, Scotland and Wales? In contrast to the correlates of linguistic distance, the big picture on the geographic structure of these dialects remains largely unchanged from the results of Szmrecsanyi (2013: 154): "a tripartite division (Scottish English dialects versus Northern English English dialects versus Southern English English dialects)". This division has emerged across virtually all cluster analyses and splits graphs in Chapter 5. As in Szmrecsanyi's analysis, the general pattern is broken by individual outliers that do not behave quite like their geographical neighbors. Two of the original outliers, Denbighshire and Warwickshire, had to be removed from the analysis due to the lack of relevant metadata, but the other four remain: Middlesex was found to fit within the general Southern English group in all model-based cluster analyses, which constitutes a departure from Szmrecsanyi's results. A similar case is the Scottish outlier Banffshire, which in general fell within the Scottish group when using normalization, but showed quite abrupt color differences in the corresponding continuum map. Both of the model-based continuum maps, however, show quite smooth color transitions here. In contrast, the Scottish outlier Dumfriesshire, which does not fall into any group in Szmrecsanyi's WPGMA analysis and is part of the main Scottish group using Ward's method, is now placed away from the Scottish Lowlands in both models. Instead, it enters either a geographically widespread group (in the lmer model predictions), or a group of dialects centered in the North of England (in the GAM predictions). The final original outlier, Leicestershire, also emerges as a measuring point on which the models differ. The next section will discuss these outliers and what features contribute to their status. I will then discuss the North of England and the areas of contention between different dialect area classifications in the literature in light of the new analyses. A discussion of the differences within the South of England will conclude this section.

6.3.1. Revisiting the outliers

Szmrecsanyi (2013: Chapter 7) provides a discussion of the features that are significantly different between the outlying measuring points and a geographically close neighbor. Let us consider these differences in light of the geographical distribution as determined by the modeling processes. For Middlesex, which was compared to London, the list consists of the three primary verbs *be*, *have*, and *do*, as well as the pronominal features *us* and *them*. The differences in primary verb usage are difficult to explain; the lmer model uncovers similar differences (see the county labels in maps 10a to 11a), but reduces their impact in absolute terms due to both the relatively sparse data from Middlesex and the overall low variability of this feature between counties. The GAMs put the results for these features in a broader perspective: *to be* and *to have* are involved in larger-scale patterns, such that *to be* increases in frequency as one moves north, and *to have* becomes less frequent as one moves northwest from Kent (with other high-frequency areas residing in the Southwest and in Yorkshire). Finally, *to do* is most frequent in the Southwest, and the particularly low frequency of this feature in Middlesex does not require larger adjustments to the general pattern. From a global perspective, therefore, these differences are not particularly relevant, and add little to the final distances. The dialectologically more immediately relevant features *them* and *us* are somewhat different. First, neither lmer- nor GAM-based distances find a particularly strong difference between London and Middlesex for *them*. Instead, the feature is comparably rare in both[1]. For *us*, we do find a notable difference, as a steep frequency cline runs right through the London-Middlesex area. The difference in this individual feature alone, however, is not strong enough to remove Middlesex from London (in the GAM) or the general Southern English group (in the lmer-based analyses).

Banffshire, an outlier with particularly low amounts of running text, exhibited a single significant difference to its rather distant neighbor Angus – again the primary verb *to be* – and two suggestive patterns, namely a low frequency of the negating suffix *-nae* and the absence of non-standard *was*. For *to be*, both models confirm that Banffshire is an atypical location. As mentioned above, the overall influence of this feature is small due to the low geographic variability. For *-nae*, both models attest its overall high frequency in Scotland, and the geographic sub-pattern in that region is of less importance. For non-standard *was*, according to Map 23a, it is actually Angus that is the outlier, while Banffshire has similar frequencies to the main Scottish group. Szmrecsanyi (2013) concludes that the status of Banffshire as an outlier is most likely a statistical aberration due to low sample size. Both model-based analyses are consistent with this and provide adjusted values.

Dumfriesshire is the other Scottish outlier, and here both models agree on its special status. This measuring point exhibits a large number of significant differences to the geographically rather close West Lothian in Szmrecsanyi's analysis, namely absence of *-nae* as well as increased frequencies of several features: the future marker *will* or *shall*, non-standard verbal *-s*, lack of inversion, *used to* as a marker of the habitual past, negative

[1] This discrepancy is due to the removal of some coding errors in the data, bringing them in line with the descriptions in Szmrecsanyi (2010a).

contraction and non-standard past tense *come*. The models agree that Dumfriesshire disprefers *going to* as a future marker compared to the other dialects in the Scottish Lowlands; it is more similar to the North of England in that regard (see Map 12a). A similar case holds for *used to*, which is frequently preferred over *would* throughout England but less so in Scotland. Map 12b illustrates the pattern. Negative contraction, here modeled in competition with auxiliary contraction, yields a similar picture. There is a large difference visible in Map 19b between the Scottish Lowlands and the North of England, indicated by the bunching of contour lines, and Dumfriesshire falls on the Northern English side of this. For non-standard verbal *-s*, it is actually West Lothian that is the outlier, as this feature is frequent both in the English Southeast and in the Scottish Lowlands, with the exception of the Lothians. Similarly for lack of inversion or auxiliaries in questions, a feature which is frequent in the southern Scottish Lowlands, but rare in West Lothian. For non-standard *come*, both models agree that the increased frequencies are particular to Dumfriesshire, although this is not visible in the GAM map due to the particularly extreme local restriction; no measuring point in the area comes even close to the prevalence of this feature in Dumfriesshire. In short, while Dumfriesshire is often similar to the other dialects in the Scottish Lowlands, for many features it actually behaves more like the North of England. The models can uncover this, and place it accordingly.

Leicestershire in the Midlands, finally, is a typical case of an outlier with very low coverage in running words. It is also a case where the models diverge, with cluster analysis performed on lmer predictions placing it in the outlier group, and GAM predictions placing it toward the Southern Englishes, together with its neighbor Nottinghamshire. When comparing it to that dialect, Szmrecsanyi found four significant differences: Leicestershire shows particularly high frequencies for non-standard *were* and non-standard past tense *done*, and particularly low frequencies for *them* and non-standard verbal *-s*. For *them*, the models disagree: while the lmer model predicts rather low frequencies, the GAM pulls it much closer to its neighbors Oxfordshire and Nottinghamshire. Overall, the results for this feature are not out of line with the frequencies in other places (thus the lower rate in the lmer models), but it does not fit the areal pattern there and is supported by little evidence. This leads the GAM to conclude that these observations are likely to be outliers. The same story holds for non-standard verbal *-s*. Considering non-standard *were*, Leicestershire shows similar frequencies to the dialects of the Southeast, with a frequency boundary running through the Midlands. There is also evidence of an east-west pattern, such that Leicestershire and Nottinghamshire exhibit higher frequencies than Shropshire toward the east. For non-standard *were*, both models find a north-south axis running through the Midlands and connecting the high-frequency areas in Somerset and Wiltshire to those in the southern part of the North of England. It is also worth noting that for many distinctly Southern features, the contour lines and county random effect BLUPs place Leicestershire with Nottinghamshire and the South. This includes the *s*-genitive (Map 4.1.1.2.2 on page 87), *ain't* (Map 18b on page 108) and invariant *don't* (Map 21b on page 119). For *got to*, multiple negation, and non-standard *was* similar circumstances apply, although the pattern in the lmer models is less clear here.

In short, the outlier status of both Middlesex and Banffshire seems to result from low textual coverage and is less pronounced in the modeled analyses, whereas Dumfriesshire is different from its Scottish neighbors in meaningful ways. Leicestershire is difficult to evaluate: the distribution for individual features is, on the whole, not too extreme, but it does not match well with the general geographic pattern in the area. Whether it constitutes a real outlier depends on how much importance the analyst, or her method, is willing to place on small amounts of data.

6.3.2. The North of England

Let us now switch the focus to the big picture, and address what light these results shine on the classification of British English dialects in general. The large-scale division that appears in all results presented here, the tripartite structure Scotland–North of England–South of England, is part of all classifications that were discussed in Section 1.2. An area of particular contention between them, however, concerned the positions of Northumberland and Lancashire. Northumberland is frequently considered to be separate from the other dialects of the North of England, for example by Goebl (2007a). On the other hand, Trudgill (1999) in his classification of modern dialects and Inoue (1996) group it with the North. Szmrecsanyi (2013) finds Northumberland to consistently fall into the Scottish cluster. In the analyses presented here, this is largely corroborated: in most maps, Northumberland joins a group with at least one dialect from the Scottish Lowlands, and usually this group is closer to the other dialects of the Scottish Lowlands than to the North of England.

It is the network diagrams in Section 5.3 that especially shed light on this: in all of them, Northumberland was found to be in a reticulation with both Scottish and (mostly) Northern English dialects.

What is the linguistic basis for this classification? Table 6.1 shows a comparison of Northumberland to the four other varieties from the North of England and to four geographically close dialects in the Scottish Lowlands: Peeblesshire, Selkirkshire and the Lothians. All values are based on lmer predictions to avoid the potential oversmoothing that may be present in the GAM values. The column *distance* shows the difference between median distances to either group; negative values indicate greater similarity to the North. Features with an absolute distance of less than 0.1 are ignored here. The remaining columns show the predicted frequencies or odds for the Scottish Lowlands, Northumberland and the North of England. For several predominantly English features, Northumberland behaves more like the dialects toward its South, these include *them* (Feature 6), multiple negation (Feature 33), *got to* (Feature 26) and lack of inversion (Feature 55), which is only frequent in one dialect from the Scottish Lowlands, Dumfriesshire. For many other features, Northumberland is closer to Scotland, especially to its closest neighbors Peeblesshire and Selkirkshire. The most important is the Scottish negating suffix *-nae* (Feature 31), which is very rare in the rest of the North of England, but appears slightly more often in Northumberland. Similar cases are unsplit *for to* (Feature 50) and infinitival complementation (Feature 51/52), which are both relatively rare in the North, but especially frequent in Northumberland and its northern neighbors. Then, there are features that are frequent in the North, but less so in Northumberland and the Scottish Lowlands. These include non-standard *were* (Feature 45), *wh*-relativization (Feature 46), and explicit complementation using *that* (Feature 53/54).

Another way to look at this distribution involves the feature clusters that were determined in Section 4.1.4.2. Four clusters were identified: the light blue cluster, which covers mostly features of the English Southeast, the dark blue cluster covering generally English features with a bias toward the Southwest, the red cluster with features of the Scottish Lowlands and, finally, the dark red cluster that is associated with features of the young dialects in the Highlands and Hebrides. Table 6.2 displays the lmer-based values for the Scottish Lowlands, the North of England, and Northumberland. For all feature clusters except the final one, Northumberland shows mean values that are intermediate between those of Scotland and the other varieties of the North. In other words, Northumberland is more English than the Scottish Lowlands, but less than the North of England; similarly it is also more Scottish than the North, but less so than Scotland. In addition, both Northumberland and Scotland exhibit slightly higher frequencies of the features that are particularly distinctive for the Southeast of England.

Northumberland is not the only dialect of the North for which the existing classifications disagree. Lancashire was found to be part of the Midlands in various schemes, such as the one based on Middle English by Baugh & Cable (1993), the traditional dialect division in Trudgill (1999), and the dialectometric analysis by Goebl (2007a). On the other hand, Trudgill's modern dialect classification considers it to be part of the North, as does Inoue (1996) from the perspective of perceptual dialectometry. Szmrecsanyi (2013) similarly finds Yorkshire to consistently cluster with the other varieties of the North, with the

Feature		distance	ScL	NBL	North
6:	*them*	−0.36	1.40	3.69	4.11
55:	lack of inversion	−0.33	0.68	2.51	1.29
33:	mult. negation	−0.27	0.74	2.37	1.86
26:	*got to*	−0.25	1.55	4.16	2.80
19/20:	habitual marking	−0.21	1.17	3.74	1.80
43:	zero aux. progressive	−0.14	0.54	0.29	0.39
4:	*ye*	−0.12	0.12	0.68	0.17
24:	*must*	−0.10	3.48	2.62	3.05
46:	*wh*-rel.	0.11	12.88	11.04	18.52
53/54:	zero/that complementation	0.12	10.03	6.84	12.92
45:	nonst. *were*	0.15	1.71	3.84	4.34
30:	nonst. *come*	0.16	1.34	1.91	1.18
39:	nonst. verbal *s*	0.16	6.45	6.24	4.04
11/12:	number + *year(s)*	0.16	0.31	1.64	0.28
51/52:	inf./ger. complementation	0.31	0.83	1.21	0.31
50:	*for to*	0.42	2.01	2.67	0.39
31:	*-nae*	0.82	11.00	3.70	0.10

Table 6.1.: Features associating Northumberland with the Scottish Lowlands or the North of England, using lmer predictions. *Distance* shows the median feature distance of Northumberland to the North of England minus the median feature distance to the Scottish Lowlands. The remaining columns show mean predicted frequencies *pttw* or odds for the Scottish Lowlands, Northumberland and the North of England.

.

Cluster	ScL	NBL	N	LAN	Mid
light blue	0.19	0.15	−0.13	0.04	0.27
dark blue	−0.70	−0.23	0.14	0.38	0.76
red	0.74	0.52	0.14	−0.03	−0.60
dark red	−0.22	−0.24	−0.20	−0.24	−0.35

Table 6.2.: Comparison of average feature bundle values in the Scottish Lowlands, the North of England and the Midlands. Lower values indicate that features in that bundle are rarer than in other varieties. For the classification of features see the cluster analysis in section 4.1.4.2

Feature		distance	North	LAN	Midlands
16:	*have got*	−0.59	0.749	0.80	4.03
51/52:	inf./ger. complementation	−0.43	0.379	0.23	1.00
40/41:	*don't/doesn't*	−0.30	0.094	0.42	1.68
26:	*got to*	−0.23	3.648	2.45	9.21
46:	*wh*-rel.	−0.19	13.761	20.57	10.27
37/38:	*wasn't/weren't*	−0.18	0.424	0.61	0.27
6:	*them*	−0.15	3.889	4.82	4.10
5:	*us*	0.11	9.521	14.65	13.36
33:	mult. negation	0.12	2.331	1.43	2.46
19/20:	habitual marking	0.12	2.141	2.29	2.48
4:	*ye*	0.14	0.168	1.40	0.18
42:	*there is*	0.17	11.090	5.57	7.70
48:	rel. *that*	0.17	19.606	10.65	13.62
34/35:	contraction with negation	0.24	7.788	2.65	6.38
30:	nonst. *come*	0.29	1.179	3.51	2.69

Table 6.3.: Features associating Northumberland with the Scottish Lowlands or the North of England, using lmer predictions. *Distance* shows the median feature distance of Lancashire to the Midlands minus the median feature distance to the North of England. The remaining columns show mean predicted frequencies *pttw* or odds for the North of England, Lancashire, and the Midlands.

exception of Northumberland as noted above. The clusters based on lmer models and especially n-gram frequencies or reliability scores, on the other hand, place Lancashire often together with parts of the Midlands or even the South. The NeighborNet analyses confirmed this, although a dual membership is only particularly notable for the lmer-based splits graph.

Table 6.3 shows a comparison of the features contributing to this, as in the corresponding table for Northumberland above. Here, Lancashire is compared to the dialects of the North and the Midlands. To keep the number of dialects in each group the same, the most centrally north dialect of the English South, Oxfordshire, was included as part of the Midlands. We find that Lancashire is more similar to the North for two features that are particularly distinctive for the Midlands. They are *have got* (Feature 16) and *got to* (Feature 26). Similarly, infinitival complementation (Features 51/52) is used more often in the Midlands, particularly in the western Midlands, but more rarely in the North except for Northumberland. For the generally southern Feature 40/41, invariant *don't* instead of *doesn't*, Lancashire is more similar to the North as well. The same is true for the probability of *wasn't* as opposed to *weren't* (Features 37/38), which is higher in the North, and for *them* (Feature 6), which is particularly frequent throughout the North, except for Westmorland.

On the other hand, Lancashire is more similar to the Midlands for other very Northern features: the relativizer *that* (Feature 48) is used less often in the Midlands and Lancashire, but frequent in the North and in Scotland, and similarly for *there is/was* with plural subjects (Feature 42). Some Southern and Midlands features are also relatively frequent in Lancashire, such as *us* (Feature 5) and *used to* as a marker of habituality (Feature 19/20). There are also two features that are relatively frequent in the Midlands, Scotland and in Lancashire, but less so in the rest of the North; these are auxiliary contraction (Features 34/35), *ye* (Feature 4), and non-standard *come* (Feature 30).

Again, we can also compare the areas using the 4 feature clusters in Table 6.2. For all of them, Lancashire shows values intermediate between the North and the Midlands. It is more Southern and less Scottish than the North, yet for all clusters it is closer to the North than to the Midlands.

The North of England has received considerable linguistic attention in recent years (e.g. Wales 2006, Montgomery 2007). Its opposition to a broad "South" is a culturally and economically important and salient distinction. That said, Montgomery (2007: 1) notes a "lack of a satisfactory definition" for the concept of the English North, and different authors place the border in very different places. For language variation, the same is true, as evidenced by the disagreements in expert classifications and in the placement of boundaries in perceptual dialectology experiments (Montgomery 2007). The investigations here and in Szmrecsanyi (2013) have found that, at least as far as morphosyntax is concerned, there is a Northern core, comprising Westmorland, Durham, and Yorkshire, with Lancashire and Northumberland somewhat removed toward either the southern dialects or Scotland. Even within that core, however, there is gradience, as the continuum maps in Section 5.3 show. This gradience runs from the North to the South, and the dialects become more Scottish as one moves toward the North. The feature clusters confirm this: no distinctively Northern English feature cluster emerges, and the North is placed intermediate for both the English and the Scottish features. That said, this is a trend in the aggregate; for individual features, the pattern may well be reversed. Particularly notable are, for instance, the sharp transition between the North and Scotland for Feature 6, *them* or *never* as a past tense negator, which is more frequent in the North than in the southern Scottish Lowlands, but less frequent in the Midlands. Finally, there are cases where the North forms a buffer zone between frequencies that are more similar in Scotland and in the Midlands. These include future markers, where the North shows a much greater preference for *will* or *shall* than either Scotland or the Midlands (Map 12a on page 94), and gerundial complementation (Map 27a on page 133). It is these features, then, that morphosyntactically distinguish the North from its neighbors the strongest.

6.3.3. The South and Midlands

There are two more points of contention between classification schemes that were discussed in Chapter 1.2. The first concerned the status of the Midlands, which were divided into either one or several groups in most schemes, but did not emerge as a coherent group in Szmrecsanyi (2013). This result is broadly confirmed here. The general tendency is for the Midlands to be split into one county that groups with Southern English, Nottinghamshire, and one county that is more similar to the North, Shropshire. Leicestershire, the third county, alternates between these two. Does this mean that the Midlands as a distinct region do not exist, at least as far as morphosyntax is concerned? The evidence seems to point in this direction. For many features that are strongly associated with Southern English, the pattern that was observed for the aggregated data holds well. These include *ain't* (Feature 32, Map 18b on page 108), invariant *don't* (Features 40/41, Map 21b on page 119), and to a slightly lower degree multiple negation (Map 19a on page 113), where the increased frequency extends to the eastern part of the English North. It should be noted, however, that FRED is relatively sparse with regard to the counties of the Midlands, and that one or several definite groups might appear if more counties were included.

Concerning the South of England, no method found a clear split between the Southeast and the Southwest, despite the fact that this division is included in most dialect area classifications. Two explanations seem plausible: first, this distinction may not be warranted on the basis of morphosyntax alone. This explanation is consistent with survey articles such as Anderwald (2008) and Wagner (2008), which note that many of the non-standard features of both the Southeast and the Southwest also appear in other regions, although "quantitative differences may be hiding behind qualitative similarities" (Anderwald 2008: 460). Only one sub-area of the Southwest consistently emerges as a group: the central western area around Somerset and Wiltshire. One could therefore conclude that while the quantitative and qualitative distribution of features across Britain does exhibit a clear pattern, it only does so for the South to a very small and localized degree.

The other possibility is that the feature catalog used here does not have sufficient power to accurately distinguish the Southwest from the Southeast. Wagner (2008: 436f.) lists the following features as uniquely Southwestern: pronoun exchange, "gendered" pronouns, unemphatic periphrastic *do* as tense carrier, and possibly mass/count distinction in demonstrative pronouns as well as otiose *of*. Only pronoun exchange and *do* are included

bigram	bigram	correlation	CON	DEV	SOM	WIL	other
PPHS2.VBM	*they 'm*	0.95	0.44	1.10	1.2	1.03	0.00
MD.VVD	*first started*	0.84	0.00	0.22	0.0	0.12	1.24
PPHS2.NNT2	*they days*	0.75	0.85	5.44	1.5	1.42	0.13
RP.NP1	*down Churchtown*	0.75	10.34	4.22	2.0	2.90	0.93
CST.VV0	*that come*	0.70	2.04	1.28	2.7	3.03	0.66
VV0.RP	*come down*	0.69	42.42	24.39	35.1	32.31	16.83
PPY.VBM	*you 'm*	0.68	0.00	0.42	2.4	0.60	0.00
RL.CCB	*there but*	0.65	2.74	2.58	2.4	2.31	1.43
DD1.PN1	*that one*	0.58	6.25	5.66	10.6	4.34	2.45
RL.VVG	*there working*	0.58	5.61	2.94	2.5	1.01	0.80

Table 6.4.: Bigrams with the highest correlation between distances in reliability scores and differences based on classification as a Southwestern dialect. Higher correlation scores indicate that this bigram is either distinctively frequent or rare in the Southwest. Remaining columns show normalized bigram frequencies for each county in the Southwest as well as the average normalized frequency in the other counties.

in the feature set, and both are in aggregate features that also cover other phenomena (possessive *us* for Feature 5, and general frequency of *to do* for Feature 13). We would then expect to find the Southwest as a more cohesive group if more of these features were included. To a small degree, the bottom-up methods confirm this: for the network based on bigram reliability scores (Figure 5.4 on page 215) we do find the West Country area as a cohesive subgroup, consisting of Cornwall, Devon, Somerset and Wiltshire. This is also consistent with the result of the bigram analysis, where forms of *do* followed by a lexical verb in the infinitive were found to be rather geographically distinctive, with frequencies centered in the Southwest.

We can test this by considering the bigrams that are most distinctive for the Southwest. To do so, the correlation coefficients between distances resulting from individual bigram reliability scores are compared against the classification into a Southwestern core (Cornwall, Devon, Somerset and Wiltshire). In other words, the distances in this matrix are zero for all comparisons either within or completely outside this group, and one for all comparisons between the groups. Again, only bigrams where at least seven pairwise combinations are significant enter consideration. The ten most distinctively Southwestern features can be found in Table 6.4. First, there are two bigrams involving extension of first person singular *am* to other persons, more specifically the third person plural (1a) and the second person (1b). Wagner (2008: 433) lists this as the traditional West

Country paradigm for *be*, noting that it can be considered antiquated and is decreasing in frequency. In the corpus, both are generally rarely used even by speakers that have this form available, but instances are completely restricted to the Southwest. The third possibility listed by Wagner, *we'm*, is also attested in the data, but fails to reach the required number of pairwise significant values as there are only three tokens in total. All of them are from Somerset, which also has the most tokens for the other two variants, indicating that non-standard *am* is most alive there. In contrast, Cornwall has the fewest tokens for *they'm*, and none at all for *you'm*.

(1) a. And they$_{\texttt{PPHS2}}$ 'm$_{\texttt{VBM}}$ still down there now, yeah. [WIL_001]
 b. You$_{\texttt{PPY}}$ 'm$_{\texttt{VBM}}$ watch what you$_{\texttt{PPY}}$ 'm$_{\texttt{VBM}}$ buying now. [SOM_005]
 c. Fortune that was in they$_{\texttt{PPHS2}}$ days$_{\texttt{NNT2}}$. [DEV_005]

Them instead of demonstrative *those* is Feature 6 in the manually selected feature set. In Section 4.2.2.3, we have seen that this is especially frequent in FRED-S after temporal nouns. Wagner (2008: 427) lists *they* as an alternative demonstrative pronoun in the Southwest, and we find precisely this combination, *they* followed by a temporal plural noun, as the third-most distinctive bigram. An example is given in (1c). This form is mostly restricted to the Southwest, although a small number of observations in Midlothian can be found. Two other attestations, from Kent and Nottinghamshire, turned out to result from disfluencies.

(2) a. And she lived up$_{\texttt{RP}}$ London$_{\texttt{NP1}}$. [WIL_008]
 b. Probably come$_{\texttt{VV0}}$ out$_{\texttt{RP}}$ churchyard. [CON_001]
 c. There 's another picture there$_{\texttt{RL}}$ but$_{\texttt{CCB}}$. . .[WIL_005]
 d. Yes , go down there$_{\texttt{RL}}$ collecting$_{\texttt{VVG}}$ cockles! [DEV_008]

Four features involve either prepositional adverbs/particles (`RP`) or locative adverbs (`RL`), two unigrams that are also among the ten most distinctive unigrams for the Southwest. Many of the former are actually prepositional usages, as in example (2a). Wagner (2008: 431) lists the use of *up* and *down* as prepositions indicating an east/west distinction as a frequent feature of the Southwest. The bigram measures agree; when a singular proper noun follows, frequencies in the Southwest are in general at least twice as high as in other counties. Following an unmarked lexical verb form, most usages occur clearly as part of particle verbs, and this is a frequent POS tag sequence in all counties. Nevertheless, it is much more frequent in the Southwest. Non-standard past tense *come*, as in example (2b) may play a role here, as standard *came* would be marked as `VVD`, the past tense of a lexical verb. The prepositional usage of up and down is likely to play a role as well. For the bigrams involving locative adverbs, a similar story holds. They are somewhat frequent throughout Britain, but particularly frequent in the Southwest. The specific bigrams are locative adverbs followed by either *but*, as in (2c), or by the *-ing* form of a lexical verb (2d). The interpretation here is not quite straight-forward, although many of the tokens again seem to involve the collocates *down there* and similar.

Two more bigrams that are particularly frequent in the Southwest are *that* (as a conjunction) followed by an unmarked lexical verb (3a) or a singular determiner (usually *that*, sometimes *this* or *another*) followed by an indefinite pronoun (3b). Many instances of the first pattern involve non-standard verb forms such as *come*, which either lack explicit past tense marking or third-person verbal *-s*. Nevertheless, both combinations are much more frequent in the Southwest, and other combinations involving *that* are not particularly rare there, indicating that a real dialectal difference may exist.[2]

(3) a. That$_{\texttt{CST}}$ come$_{\texttt{VV0}}$ from the hill up on the top [...] [SOM_006]
b. There look, that$_{\texttt{DD1}}$ one$_{\texttt{PN1}}$ there. [WIL_005]
c. I first$_{\texttt{MD}}$ smoked$_{\texttt{VVD}}$ before the War. [WIL_008]

Finally, while all bigrams discussed so far are particularly frequent in the Southwest, the second-most distinctive bigram is actually an absence. It involves an ordinal number, almost always *first*, followed by the past tense of a lexical verb, as in (3c). This sequence is moderately frequent throughout all counties, but very rare in the Southwest, with no attestations in Cornwall and Somerset at all. The absence of this pattern is also restricted to the informants, as the interviewers did use it.

Other relevant bigrams, such as the aforementioned *do*-related patterns or cases of pronoun exchange (e. g. *one of they*), can be found slightly lower on the list. In other words, looking at bigrams, important features of the dialect grammar of the Southwest can be found. Looking at the lmer or GAM model predictions, however, almost no feature shows particular absence or presence in the Southwest. Most features exhibit low correlation values and only atypical features score highly, such as the rare and only weakly geographically distributed Feature 10, synthetic adjective comparison.

Overall, therefore, while there is evidence for an east/west split in the South of England, it is much less important than the larger pattern, contrasting the North and the South. Furthermore, even for those n-grams where a regional pattern is evident, in terms of frequencies the signal is sometimes less clear. For most of the bigrams in Table 6.4,

[2] Trudgill (2009a: 106) reports that in Norfolk in East Anglia, *that* can function as the third person singular neuter personal pronoun. Distinguishing such uses from demonstrative ones can be difficult, and in most instances of the pattern *that* is used as a relative marker.

one county usually shows a normalized frequency that does not quite fit the others, and frequencies tend to be rather low. Kortmann & Wagner (2010: 284) note that "high text frequency is not a necessary prerequisite for salience", and the results presented here seem to agree with that. A rather salient distinction in British English dialects, that between the Southeast and the Southwest of England, only emerges when considering rather rare morphosyntactic features. Emphasizing rare phenomena is, of course, an approach that has been successfully used in dialectometric analysis. It is the central idea behind the *Gewichtender Identitätswert* introduced by Goebl (1984), which has performed admirably in empirical analysis (Nerbonne & Kleiweg 2007). Szmrecsanyi (2013: 25) also lists such emphasis as one of the reasons for using a logarithmic transformation on the normalized frequency counts. This provides a challenge to the CBDM enterprise: rare features are precisely where the corpus-based approach fares worst, as discussed in Szmrecsanyi (2013: 37f.). For them, absence is particularly likely to result from chance, and therefore the results will be subject to high amounts of noise. This was also shown via simulation in Section 3.2.2. Using n-grams does not solve this problem - absences as well as comparably high values may still result from chance. The pure amount of features that enter consideration in the bottom-up analysis, however, appears to remedy such problems at least partially, and allows finer subgroups to emerge.

6.4. Outlook and concluding remarks

Who is the corpus-based dialectometrist to trust? I have argued that there are at least three positions she could take.

First, the analyst could place her trust completely in her data. This view has many advantages; crucially, it is closest to the actual observations, and keeps all of the noisiness that is "part of linguistic reality" (Szmrecsanyi 2013: 163). On the other hand, such an approach is likely to be led astray by this noise. Naturalistic corpus data is influenced by more factors than geography alone, and data sparsity is, at least at present, an issue that is endemic to dialect corpora. The analyst may therefore end up measuring not the linguistic diversity, but incidental factors such as subcorpus size.

The next option is to trust reliable data more than less reliable data. The intuition here is simple: if a certain dialect has extreme frequency distributions that are supported by small quantities of data, these distributions should count less than well-supported ones, as they are subject to higher variance. Only reliable differences should have a strong effect on the result. As was shown in the simulation experiment in Section 3.2.2.1, doing so can increase the accuracy of the measurements for less frequent features in particular. This, however, requires the analyst to accept two side-effects: first, some of the true variability in the data may be lost. Second, this method may overestimate the differences where a lot of data is available.

The third option is to let geography guide the analysis. Here, individual observations are considered in the context of their spatial neighbors and the overall shape of the feature distribution; if a small number of speakers has an undue influence, they are smoothed toward the overall pattern. Sharp transitions then require solid support, while gradual transitions that make sense in the big picture are allowed more freely. The downside to this is that it is the strongest assumption, and abstracts away the most from the data. It leads to the most consistent result geolinguistically, but this is hardly surprising as the method assumes this outcome.

No approach is strictly superior to the others. What does this mean for the CBDM enterprise? First, the best way to deal with sparsity is, of course, to have data that are plentiful enough. With current dialect corpora, this is unfortunately not an option, because dialectologically suitable texts are not easy to come by. More oral history interviews could be included in FRED, but preparing them for linguistic research is a labor-intensive task. For modern dialectal variation, deriving data from internet resources is an exciting possibility (Ruette 2011, Grieve et al. 2013). This, however, largely limits the applicability of the methods to dialectal variation that is still in use by modern, computer-literate speakers, and therefore of questionable use for traditional dialectology. For those purposes, and when compiling larger corpora is not feasible, having a synoptic view of the different methods seems to be the best choice. Instead of asking "What is the big picture of morphosyntactic variation in Britain?", the question then becomes "What is morphosyntactic variation in Britain like given certain, specific assumptions?" The researcher can then investigate what remains similar (here, among others, the large-scale split into Scotland, the North and the South of England) versus what is different (here, the details of the splits and outliers, and the general relation between linguistic distances and operationalizations of geography).

One area where the model-based variants have a clear advantage, however, is that they can cover the full spectrum from the perspective of the *jeweler's eye* to that of the *bird's eye*. Given the successes of the VARBRUL approach and its modern descendants in sociolinguistics and the recent surge of interest in probabilistic grammar, regression modeling is likely to be an attractive technique for dialectologists taking a single-feature perspective. Depending on the precise research interests, both lmer models and GAMs seem suitable for this task. When several of such studies use the same corpus as their data source, they can be combined into a more precise picture using the dialectometric tool set as an additional bonus.

Let me conclude by sketching directions in which the methods proposed here could be developed. For the top-down approach, I see two major areas: including more linguistic detail into the models, and extending the approach to data with different properties, especially to data from linguistic levels other than morphosyntax. The first involves annotating the individual tokens for contextual factors, and including them in the models as predictors, similar to the case study presented in Section 4.1.2. For example, the genitive tokens collected for Features 8/9 could be annotated for animacy, definiteness and length of the constituents using a coding scheme like the one used in Wolk et al. (2013). Not only could this lead to interesting results regarding the genitive alternation, it should also strengthen the geographic signal, as this process should make the locations more comparable. An especially interesting possibility involves models that allow the effect of these predictors to vary based on geography. For example, in some dialects the genitive choice might be determined by animacy more strongly than in others, similar to how different varieties of English seem to be influenced by end weight in slightly different ways (Bresnan & Ford 2010, Wolk et al. 2013). Whether current dialect corpora are large enough to handle such small effects remains to be seen. More extensive models also raise the question of how to best aggregate over them. Should only one combination of features count, as in this study, where only the predictions for the default speaker (male, mean age) are included? How should it be chosen? Or should the aggregation be based on several predictors or their combinations? Should they be averaged before aggregating or count as single features? Should this averaging, if any, be weighted by the frequency of that feature combination, or should each count the same? While data sparsity and annotations are likely to be the greatest challenges of this project, the aggregation step is also not completely straightforward.

The second direction concerns extensions to different data types. In this work, only frequencies and binary alternations were analyzed, which suffices for many morphosyntactic features. However, Szmrecsanyi (2013) argues that the CBDM approach is relevant to all linguistic levels, and especially to the mainstays of dialectometry, lexical and phonetic/phonological variation. And here, the limitation to frequencies and binary alternations is especially troubling. Consider the data set in Streck (2012): a corpus of interviews with southwest German dialect speakers was searched for 172 lexemes instantiating 38 phonological variables. The results of this search were then coded for how the variables were realized, and Levenshtein distance was applied to them. In some ways, the results obtained by Streck (2012) mirror Szmrecsanyi's: interpretable larger areas with frequent outliers, and an overall rather low correlation between geographic and linguistic distance.

Some lexemes appear frequently while others are only used in few interviews; similarly, the number of observations varies heavily by location. Therefore, it is possible that the distances are influenced by data sparsity, and the preliminary results of an ongoing collaboration with Tobias Streck suggest this to be the case. A GAM-based strategy could be employed to reduce the influence of this factor. However, many of the lexemes have more than two realizations, with some (such as *zwei* 'two' having over 10). While it would be possible to simply model the frequencies of individual realizations, this is profoundly unsatisfying: consider a hypothetical case where location L1 has 10 tokens of realization A and location L2 has 5; neither has any other realization attested in the data set. Location L3, however, has 10 tokens of A and 5 of C. Going simply by frequencies, L2 and L3 seem to be equidistant from L1. However, L1 and L2 have the same observed probability of choosing realization A, only the base rate of the lexeme is different. A proper metric should place L1 much closer to L2 than to L3. Multinomial models are an extension to logistic regression that can represent alternations with more than two realizations, but is computationally difficult and neither `lmer` nor the GAM implementation `mgcv` provide adequate support. Furthermore, some realizations are very rare and may only appear once. The model may well assign a very low probability for this realization to occur in any location. But we know that it did occur, and it seems important to keep this represented in the data at least to some degree. Both issues necessitate extensions to the method presented here.

Finally, let us turn to the bottom-up approach. Here, the opportunities for follow-up research are endless. Most urgent is replication of the method on different data sets, as the reliability measure is so new and might, despite its performance on FRED-S, not be reliable itself. Replication and development on a large corpus would be best. Unfortunately dialect corpora larger than FRED-S are hard to come by, but there is material for international varieties of English: several components of the *International Corpus of English* (ICE, Greenbaum & Nelson 1996) have recently become available in versions that are POS-annotated using CLAWS7. I have already begun to apply the method on this data set, and the early results look promising. The larger size of the ICE components – one million words each – will also allow the study of trigrams and possibly even quadgrams. This calls for a method to integrate the different lengths, especially for feature identification. Optimally, patterns only show up where they are most relevant, so that the core patterns can be determined more easily. A way to integrate regional variation with age, gender and (for ICE) register variation would also be very welcome. Experimentation with different tag sets (POS or semantic, e.g. USAS (Rayson et al. 2004), also available for ICE), untagged data (in the spirit of Gries & Mukherjee (2010)), hybrid POS/lexical/semantic n-grams, and different normalization methods may also prove fruitful. If this research confirms the effectiveness of reliability as a measure, it will be important to find a convincing linguistic or cognitive motivation for this. As it stands, the face validity of this measure is rather low – it is certainly not immediately obvious that permuting corpora in this way will lead to meaningful patterns.

As an afterword, let me quote Ihalainen (1988: 581), who in a footnote states:

> The problem of obtaining good data for syntactic analysis has worried me since my first paper on dialectal syntax [. . .], and it still does [. . .]. We are very far from the day when one feels comfortable about syntactic data. [references omitted]

25 years have passed since the publication of this article, and what he called for in his article is reality now: we have large data sets such as FRED, and they exist in tagged versions that make them easy to search. Nevertheless, the availability of good data remains an issue. We have certainly come a long way, but we should not feel too comfortable yet.

A. CLAWS7 Tag Set[1]

APPGE – possessive pronoun, pre-nominal (e.g. *my, your, our*)

AT – article (e.g. *the, no*)

AT1 – singular article (e.g. *a, an, every*)

BCL – before-clause marker (e.g. *in order (that),in order (to)*)

CC – coordinating conjunction (e.g. *and, or*)

CCB – adversative coordinating conjunction (*but*)

CS – subordinating conjunction (e.g. *if, because, unless, so, for*)

CSA – *as* (as conjunction)

CSN – *than* (as conjunction)

CST – *that* (as conjunction)

CSW – *whether* (as conjunction)

DA – after-determiner or post-determiner capable of pronominal function (e.g. *such, former, same*)

DA1 – singular after-determiner (e.g. *little, much*)

DA2 – plural after-determiner (e.g. *few, several, many*)

DAR – comparative after-determiner (e.g. *more, less, fewer*)

DAT – superlative after-determiner (e.g. *most, least, fewest*)

DB – before determiner or pre-determiner capable of pronominal function (*all, half)*

DB2 – plural before-determiner (*both*)

DD – determiner (capable of pronominal function) (e.g. *any, some*)

DD1 – singular determiner (e.g. *this, that, another*)

[1] Based on the list available at http://ucrel.lancs.ac.uk/claws7tags.html

DD2	plural determiner (*these,those*)
DDQ	*wh*-determiner (*which, what*)
DDQGE	*wh*-determiner, genitive (*whose*)
DDQV	*wh*-ever determiner, (*whichever, whatever*)
EX	existential *there*
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (*'* or *'s*)
IF	*for* (as preposition)
II	general preposition
IO	*of* (as preposition)
IW	*with, without* (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. *older, better, stronger*)
JJT	general superlative adjective (e.g. *oldest, best, strongest*)
JK	catenative adjective (*able* in *be able to*, *willing* in *be willing to*)
MC	cardinal number,neutral for number (*two, three...*)
MC1	singular cardinal number (*one*)
MC2	plural cardinal number (e.g. *sixes, sevens*)
MCGE	genitive cardinal number, neutral for number (*two's, 100's*)
MCMC	hyphenated number (*40-50*, *1770-1827*)
MD	ordinal number (e.g. *first, second, next, last*)
MF	fraction,neutral for number (e.g. *quarters, two-thirds*)
ND1	singular noun of direction (e.g. *north, southeast*)
NN	common noun, neutral for number (e.g. *sheep, cod, headquarters*)
NN1	singular common noun (e.g. *book, girl*)

NN2 – plural common noun (e.g. *books, girls*)

NNA – following noun of title (e.g. *M.A.*)

NNB – preceding noun of title (e.g. *Mr., Prof.*)

NNL1 – singular locative noun (e.g. *Island, Street*)

NNL2 – plural locative noun (e.g. *Islands, Streets*)

NNO – numeral noun, neutral for number (e.g. *dozen, hundred*)

NNO2 – numeral noun, plural (e.g. *hundreds, thousands*)

NNT1 – temporal noun, singular (e.g. *day, week, year*)

NNT2 – temporal noun, plural (e.g. *days, weeks, years*)

NNU – unit of measurement, neutral for number (e.g. *in, cc*)

NNU1 – singular unit of measurement (e.g. *inch, centimetre*)

NNU2 – plural unit of measurement (e.g. *ins., feet*)

NP – proper noun, neutral for number (e.g. *IBM, Andes*)

NP1 – singular proper noun (e.g. *London, Jane, Frederick*)

NP2 – plural proper noun (e.g. *Browns, Reagans, Koreas*)

NPD1 – singular weekday noun (e.g. *Sunday*)

NPD2 – plural weekday noun (e.g. *Sundays*)

NPM1 – singular month noun (e.g. *October*)

NPM2 – plural month noun (e.g. *Octobers*)

PN – indefinite pronoun, neutral for number (*none*)

PN1 – indefinite pronoun, singular (e.g. *anyone, everything, nobody, one*)

PNQO – objective wh-pronoun (*whom*)

PNQS – subjective wh-pronoun (*who*)

PNQV – wh-ever pronoun (*whoever*)

PNX1 – reflexive indefinite pronoun (*oneself*)

PPGE – nominal possessive personal pronoun (e.g. *mine, yours*)

PPH1 – 3rd person sing. neuter personal pronoun (*it*)

PPHO1 – 3rd person sing. objective personal pronoun (*him, her*)

PPHO2 – 3rd person plural objective personal pronoun (*them*)

PPHS1 – 3rd person sing. subjective personal pronoun (*he, she*)

PPHS2 – 3rd person plural subjective personal pronoun (*they*)

PPIO1 – 1st person sing. objective personal pronoun (*me*)

PPIO2 – 1st person plural objective personal pronoun (*us*)

PPIS1 – 1st person sing. subjective personal pronoun (*I*)

PPIS2 – 1st person plural subjective personal pronoun (*we*)

PPX1 – singular reflexive personal pronoun (e.g. *yourself, itself*)

PPX2 – plural reflexive personal pronoun (e.g. *yourselves, themselves*)

PPY – 2nd person personal pronoun (*you*)

RA – adverb, after nominal head (e.g. *else, galore*)

REX – adverb introducing appositional constructions (*namely, e.g.*)

RG – degree adverb (*very, so, too)*

RGQ – *wh*-degree adverb (*how*)

RGQV – *wh*-ever degree adverb (*however*)

RGR – comparative degree adverb (*more, less*)

RGT – superlative degree adverb (*most, least*)

RL – locative adverb (e.g. *alongside, forward*)

RP – prep. adverb, particle (e.g *about, in*)

RPK – prep. adv., catenative (*about* in *be about to*)

RR – general adverb

RRQ – *wh*-general adverb (*where, when, why, how*)

RRQV – *wh*-ever general adverb (*wherever, whenever*)

RRR – comparative general adverb (e.g. *better, longer*)

RRT – superlative general adverb (e.g. *best, longest*)

RT – quasi-nominal adverb of time (e.g. *now, tomorrow*)

TO	– infinitive marker (*to*)
UH	– interjection (e.g. *oh, yes, um*)
VB0	– *be*, base form (finite i.e. imperative, subjunctive)
VBDR	– *were*
VBDZ	– *was*
VBG	– *being*
VBI	– *be*, infinitive (*To be or not...It will be ..*)
VBM	– *am*
VBN	– *been*
VBR	– *are*
VBZ	– *is*
VD0	– *do*, base form (finite)
VDD	– *did*
VDG	– *doing*
VDI	– *do*, infinitive (*I may do...To do...*)
VDN	– *done*
VDZ	– *does*
VH0	– *have*, base form (finite)
VHD	– *had* (past tense)
VHG	– *having*
VHI	– *have*, infinitive
VHN	– *had* (past participle)
VHZ	– *has*
VM	– modal auxiliary (*can, will, would*, etc.)
VMK	– modal catenative (*ought, used*)
VV0	– base form of lexical verb (e.g. *give, work*)
VVD	– past tense of lexical verb (e.g. *gave, worked*)

VVG	– *-ing* participle of lexical verb (e.g. giving, working)
VVGK	– *-ing* participle catenative (*going* in *be going to*)
VVI	– infinitive (e.g. *to give. . .It will work. . .*)
VVN	– past participle of lexical verb (e.g. *given, worked*)
VVNK	– past participle catenative (e.g. *bound* in *be bound to*)
VVZ	– -*s* form of lexical verb (e.g. *gives, works*)
XX	– *not, n't*
ZZ1	– singular letter of the alphabet (e.g. *A,b*)
ZZ2	– plural letter of the alphabet (e.g. *A's, b's*)

B. Technical notes

The following explains how to implement the analyses described in Section 3.2. Future improvements to the material presented here, as well as the data sets or distance matrices used in this analysis, will be made available online at `http://wolki.org`,

B.1. Top-down

B.1.1. Data

I will assume a data set `FRED`, which has one row per speaker and the following columns:

- `speaker` containing the speaker's unique identifier
- `county`, containing the speaker's county
- `cAge`, containing the speaker's age (centered on zero)
- `Sex`, containing the speaker's gender
- `latitude`, containing the latitude of the interview location
- `longitude`, containing the longitude of the speaker location
- `feat1`–`feat57`, containing the raw frequency counts for each feature
- `feat1.no.words`–`feat57.no.words`, containing the number of words examined for each feature

B.1.2. lmer models

To calculate count-based models, use the number of words as an offset.

```
feat3.lmer <- glmer(feat3 ~ offset(log(feat3.no.words)) +
                       cAge * Sex + (1|county), data=FRED,
                       family="poisson")
```

To fit logistic regression models from the counts for the individual realizations, use `cbind`. The first argument is the predicted realization.

```
feat1_2.lmer <- glmer(cbind(feat2, feat1) ~ cAge * Sex +
                      (1|county), data=FRED,
                         family="binomial")
```

To extract the predicted counts or proportions, use:

```
predicted.feat3.lmer <-
   coef(feat3.lmer)[["county"]][,"(Intercept)"]
predicted.feat1_2.lmer <-
   coef(feat1_2.lmer)[["county"]][,"(Intercept)"]
```

B.1.3. GAMs

To calculate count-based models, give the number of words as an offset and specify the parameter range for the negative binomial family. After fitting the model, check the parameter for the negative binomial distribution. If it near the maximum of the range, refit the model with a higher range.

```
library(lme4)
feat3.gam <- gam(feat3 ~ offset(log(feat3.no.words)) + Sex *
   cAge +
            s(longitude, latitude), data = FRED,
            family=negbin(1:5))
```

Fitting logistic models is straightforward.

```
library(mgcv)
feat1_2.gam <- gam(cbind(feat2, feat1) ~ Sex * cAge +
                   s(longitude, latitude), data=FRED,
                      family=binomial)
```

For getting the GAM predictions, a data frame `new_data` for default speakers is necessary. It should have one row per county, with all number of words columns being set to 10,000, `Sex` being male, `cAge` being zero, and the `longitude` and `latitude` columns giving the value for the county centers.

```
predicted.feat3.gam <- predict(feat3.gam, new_data,
   type="response",
                            terms="s(longitude,latitude)"))
predicted.feat1_2.gam <- predict(feat1_2.gam, new_data,
   type="response",
                            terms="s(longitude,latitude)"))
```

B.1.4. Aggregation

Next, the values need to be transformed according to the CBDM parameters.

```
predicted.feat3.lmer <- log10(exp(predicted.feat3.lmer))
predicted.feat3.gam  <- log10(exp(predicted.feat3.lmer))

predicted.feat1_2.lmer <- log10(exp(predicted.feat1_2.lmer))
predicted.feat1_2.gam  <- log10(exp(predicted.feat1_2.lmer))
predicted.feat3.lmer[predicted.feat3.lmer < -1] <- -1
predicted.feat3.gam[predicted.feat3.gam < -1] <- -1
```

For alternations, we constrain the values to the range $[-2, 2]$

```
predicted.feat1_2.lmer[predicted.feat1_2.lmer < -2] <- -2
predicted.feat1_2.gam [predicted.feat1_2.gam  < -2] <- -2

predicted.feat1_2.lmer[predicted.feat1_2.lmer > 2] <- 2
predicted.feat1_2.gam [predicted.feat1_2.gam  > 2] <- 2
```

The results can then be combined into matrices `predicted.lmer` and `predicted.gam` using `cbind()`. The distances can then be calculated using `dist()`, and be passed to *RuG/L04* using its R interface.

B.2. Bottom-up

I am currently reworking the code for the bottom-up analysis. This version is not particularly efficient, but relatively fast and easy to understand.

Two data frames are necessary. One, `ngrams`, should one row per n-gram token in the corpus and a unique id for the desired resampling level (here conversational turns); the other `countymap`, should contain one row per resampling level id (here, the turn id) and the corresponding analysis level (here, the county). The column names for the resampling level ids must be the same in both data frames.

B.2.1. Normalization

The following code can be used to normalize a matrix with counts, where the rows are n-grams and the columns counties. All other columns have to be removed from the data before

```
normalize1 <- function(m) {
  m <- t(m)
  mnorm <- m / rowSums(m)
  result <- t(t(mnorm) * colSums(m) / colSums(mnorm))
  result[is.nan(result)] <- 0
  return(t(result))
}

normalize2 <- function(m) {
 result <- (m * ncol(m) * nrow(m)) / sum(m)
 return(result)
}


normalize <- function(df, repet=5) {
  df <- df[, colSums(df) > 0]
  normalized <- normalize1(df)
  if (repet > 1)
    for (i in 1:(repet-1))
      normalized <- normalize1(normalized)
  return(normalize2(normalized))
}
```

B.2.2. Counting and permutation

The following functions can be used to count the ngrams, pairwise (`runPair()`) or in total (`runAll()`). The parameters are as follows:

- `base`: the data frame containing the ngrams
- `perm`: the data frame mapping the basis of permutation to the groups
- `which`: the name of the column containing the groups
- `permute`: `FALSE` to get the counts for the original corpus, a number to get a list containing that many permuted counts.
- `which2`: the column name of the resampling level, only necessary for `runPair()`.

```
runAll  <- function (base, perm, which="county", permute=1) {

  doOnce <- function(base,perm,which) {
    left_join(base, perm) %>% function(x)
      eval(substitute(group_by(x, which, ngram),
                      list(which = as.name(which)))) %>%
      summarize(count=n()) %>%
      dcast(formula(sprintf("ngram~%s", which)), fill=0,
         value.var="count")
  }


  if (permute){
    lapply(1:permute, function(x){
      perm2 <- perm
      perm2[,which] <- sample(perm2[,which])
      doOnce(base,perm2,which)
    })
  }
  else
    doOnce(base,perm,which)
}
```

```
runPair  <- function(base, perm, a, b, which="county",
   which2="turn", permute=1) {

  perm2 <- perm [ perm[,which] %in% c(a,b),]
  base2 <- base [ base[,which2] %in% perm2[,which2], ]

  runAll(base2, perm2, which=which, permute=permute)

}
```

B.2.3. Analysis

To do a whole-corpus comparison, the following code can be used. Distance matrix calculation and exporting to *Rug/L04* can then happen as usual.

```
compareRuns <- function(x, original) {

  ((normalize(original[, -1]) < normalize(x[,-1])) * 1 +
      (normalize(original[, -1]) == normalize(x[,-1])) * 0.5)
}

library(dplyr)
library(reshape2)
library(magrittr)

nruns=1000

relscores <- Reduce("+", lapply(runAll(ngrams, countymap,
   permute=nruns),
                           compareRuns,
                           runAll(ngrams, countymap,
                              permute=FALSE)))/nruns

r.distinct <- 2*pmin(abs(relscores), abs(1 - relscores))
```

Finally, the following code can be used to run a pairwise analysis. `rp` is a list of significance matrices, one for each bigram.

```
runPairwise  <- function(repetitions=100, base, perm,
   which="county", which2="turn") {
  sigmatlist <- list()
  CORPORA <- unique(perm[,which])
  pairdiff <- function(a, b) pmax(a,b) - pmin(a,b)

  for (i in levels(ngrams$ngram))
    sigmatlist[[i]] <- getempty()

  for (i in 1:(length(CORPORA)-1))
    for (j in (i+1):length(CORPORA)) {

      print(paste(CORPORA[i], CORPORA[j]))

      base <- runPair(corpbase, map, CORPORA[i],
                      CORPORA[j], permute=FALSE)

      rownames(base) <- base[,1]
      base <- normalize(t(base[,-1]))

      rec.base <- pairdiff(base[,1], base[,2])

      new.list <- runPair(corpbase, map, CORPORA[i],
                          CORPORA[j], which=which,
                             which2=which2,
                             permute=repetitions)

      new.list <- lapply(new.list, function(new){
        rownames(new) <- new[,1]
        new <- normalize(t(new[,-1]))
        pairdiff(new[,1], new[,2])
      })

      sigs <- Reduce("+", lapply(new.list, function(x) x >=
         rec.base))/nrepet

      for (bigr in colnames(base)) {
        sigmatlist[[bigr]][i,j] <- sigs[bigr]
        sigmatlist[[bigr]][j,i] <- sigs[bigr]
      }
    }

  sigmatlist

}

rp <- runPairwise(repetitions=1000, base=ngrams,
   perm=countymap)
rp <- lapply(rp, function(x) {x[is.na(x)] <- 1; x})
```

```
p.distinct <- sapply(rp, function(x) sum(as.dist(x) < 0.05))
```

Bibliography

Abdi, Hervé & Dominique Valentin. 2007. Multiple correspondence analysis. In N.J. Salkind (ed.), *Encyclopedia of measurement and statistics*, 651–657. Thousand Oaks (CA): Sage.

Adger, David & Graeme Trousdale. 2007. Variation in English syntax: Theoretical implications. *English Language and Linguistics* 11(2). 261–278.

Albu, Mihai. 2006. *Quantitative analyses of typological data.* Leipzig: University of Leipzig dissertation.

Altmann, Eduardo G., Janet B. Pierrehumbert & Adilson E. Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One* 4(11). e7678.

Anderson, Peter M. 1987. *A Structural Atlas of the English Dialects.* London, New York: Croom Helm.

Anderwald, Lieselotte. 2002. *Negation in non-standard British English: Gaps, regularizations, asymmetries* (Studies in Germanic Linguistics 8). London, New York: Routledge.

Anderwald, Lieselotte. 2003. Non-standard English and typological principles: The case of negation. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of linguistic variation* (Topics in English Linguistics 43), 507–529. Berlin: Mouton de Gruyter.

Anderwald, Lieselotte. 2005. Negative concord in British English dialects. In Iyeiri (2005), 113–137.

Anderwald, Lieselotte. 2008. English in the Southeast of England: Morphology and syntax. In Kortmann & Upton (2008), 440–462.

Anderwald, Lieselotte. 2009. *The morphology of English dialects: Verb-formation in non-standard English.* Cambridge: Cambridge University Press.

Anderwald, Lieselotte & Benedikt Szmrecsanyi. 2009. Corpus linguistics and dialectology. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, vol. 2 (29 Handbücher zur Sprach- und Kommunikationswissenschaft), 1126–1140. Berlin, New York: Mouton de Gruyter.

Auer, Peter, Peter Baumann & Christian Schwarz. 2011. Vertical vs. horizontal change in the traditional dialects of southwest Germany: A quantitative approach. *taal & tongval* 63(1). 13–41.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press.

Baayen, R. Harald, Doug J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Baugh, Albert C. & Thomas Cable. 1993. *A history of the English language.* Englewood Cliffs, N.J.: Prentice-Hall 4th edn.

Beal, Joan C. & Karen P. Corrigan. 2005. *No, Nay Never*: Negation in tyneside English. In Iyeiri (2005), 139–156.

Bickel, Balthasar. 2012. Exploring similarities: p phylogenetic methods beyond phylogeny. Presented at the workshop "Phylomemetic and phylogenetic approaches in the humanities", Saturday, Nov. 24th 2012, University of Bern, Switzerland.

Bloomfield, Leonard. 1933. *Language.* New York: Holt, Rhinehart and Winston.

Bookstein, Abraham & Donald H. Kraft. 1977. Operations research applied to document indexing and retrieval decisions. *Journal of the Association for Computing Machinery* 24(3). 418–427.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Boume, Irene Kraemer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 186–213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245–259.

Bridge, Martin. 2006. "It's not always like this!" - Examining the UK's negative/auxiliary-contraction strategies. Poster presented at the NWAV 35, Columbus, OH.

Britain, David. 2010. Grammatical variation in the contemporary spoken English of England. In Andy Kirkpatrick (ed.), *The handbook of World Englishes*, 37–58. London: Routledge.

Bryant, David & Vincent Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255–65.

Chambers, J. K. 2003. *Sociolinguistic theory: Linguistic variation and its social significance.* Oxford: Blackwell.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology.* Cambridge, England: Cambridge University Press 2nd edn.

Cheshire, Jenny. 1989. A survey of dialect grammar in British English. In Michel H. A. Blanc & Josiane F. Hamers (eds.), *Problèmes théoriques et méthodologiques dans l'étude des langues/dialectes en contact aux niveaux macrologiques et micrologiques*, 50–58. Québec: Centre international de recherche sur le bilinguisme.

Cheshire, Jenny. 1998. English negation from an interactional perspective. In Ingrid Tieken-Boon van Ostade, Gunnel Tottie & Wim van der Wurff (eds.), *Negation in the history of English* (Topics in English Linguistics 26), 29–54. Berlin: Mouton de Gruyter.

Cheshire, Jenny, Viv Edwards & Pamela Whittle. 1995. Urban British dialect grammar: The question of dialect levelling. In Iwar Werlen (ed.), *Verbale Kommunikation in der Stadt*, 67–109. Tübingen: Gunter Narr Verlag. Reprinted from English World-Wide 10 (2): 185-225.

Cheshire, Jenny & Sue Fox. 2009. *Was/were* variation: A perspective from London. *Language Variation and Change* 21(1). 1–38.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12. 335–359.

Cliff, Andrew D. & J. K. Ord. 1973. *Spatial autocorrelation.* London: Pion.

Close, Joanne & Bas Aarts. 2010. Current change in the modal system of English a case study of *must*, *have to* and *have got to.* In Ursula Lenker, Judith Huber & Robert

Mailhammer (eds.), *English historical linguistics 2008: Selected papers from the fifteenth International Conference on English Historical Linguistics (ICEHL 15), Munich, 24-30 August 2008. Volume I: The history of English verbal and nominal constructions*, 65–182. Amsterdam: John Benjamins.

Cysouw, Michael. 2007. New approaches to cluster analysis of typological indices. In Reinhard Köhler & Peter Grzbek (eds.), *Exact methods in the study of language and text*, 61–76. Berlin: Mouton de Gruyter.

Davies, Mark. 2008-. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at `http://corpus.byu.edu/coca/`.

Dawson, Christina. 2011. *Shared alternatives to subject-verb agreement: The 3rd singular verb and its uses in English and Brittonic.* Freiburg: Albert-Ludwigs-Universität Freiburg dissertation.

Denison, David. 1993. *English historical syntax: Verbal constructions.* London: Longman.

Dress, Andreas W. M. & Daniel H. Huson. 2004. Constructing splits graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1. 109–115.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink & Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84(4). 710–759.

Dyen, Isidore, Joseph Kruskal & Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5).

Ellis, Alexander J. 1889. *The existing phonology of English dialects compared with that of West Saxon speech* (On Early English Pronunciation V). London: Trübner & co.

The Helsinki Corpus of British English Dialects. 2006. Department of Modern Languages, University of Helsinki.

Evans, William. 1969. 'You' and 'thou' in Northern England. *South Atlantic Bulletin* 34(4). 17–21.

Filppula, Markku, Juhani Klemola & Heli Paulasto (eds.). 2009. *Vernacular universals and language contacts.* Abingdon: Routledge.

Fischer, Olga & Wim van der Wurff. 2006. Syntax. In Richard Hogg & David Denison (eds.), *A history of the English language*, 109–198. Cambridge: Cambridge University Press.

Garside, Roger & Nicholas Smith. 1997. A hybrid grammatical tagger: Claws4. In Roger Garside, Geoffrey Leech & Tony McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora.*, 102–121. London: Longman.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Goebl, Hans. 1982. Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Denkschriften der Österreichischen Akademie der Wissenschaften, phil.-hist. Klasse, Band 157.

Goebl, Hans. 1984. *Dialektometrische Studien anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF.* Tübingen: Max Niemeyer.

Goebl, Hans. 2005. Dialektometrie. In Richard Köhler, Gabriel Altmann & Raimund G. Piotrowski (eds.), *Quantitative linguistics. An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 27), 498–531. Berlin, New York: de Gruyter.

Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4). 411–435.

Goebl, Hans. 2007a. A bunch of dialectometric flowers: a brief introduction to dialectometry. In Ute Smit, Stefan Dollinger, Julia Hüttner, Günther Kaltenböck & Ursula Lutzky (eds.), *Tracing English through time: Explorations in language variation*, 133–172. Wien: Braumüller.

Goebl, Hans. 2007b. Dialektometrische Streifzüge durch das Netz des Sprachatlasses AIS. *Ladinia* XXXI. 187–271.

Goebl, Hans. 2010. Dialectometry and quantitative mapping. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and space. An international handbook of linguistic variation, vol. 2: Language mapping* (Handbücher zur Sprach- und Kommunikationswissenschaft 30), 433–464. Berlin: Mouton.

Goebl, Hans & Guillaume Schiltz. 1997. A dialectometrical compilation of CLAE I and CLAE II. Isoglosses and dialect integration. In Wolfgang Viereck & Heinrich Ramisch (eds.), *Computer Developed Linguistic Atlas of England (CLAE)*, vol. 2, 13–21. Tübingen: Niemeyer.

Grafmiller, Jason. forthcoming. Variation in English genitives across modality and genre. English Language & Linguistics.

Greenbaum, Sidney & Gerald Nelson. 1996. The International Corpus of English (ICE) project. *World Englishes* 15(1). 3–15.

Gries, Stefan Th. & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15(4). 520–548.

Grieve, Jack. 2009. *A corpus-based regional dialect survey of grammatical variation in written Standard American English.* Flagstaff: Northern Arizona University dissertation.

Grieve, Jack, Costanza Asnaghi & Tom Ruette. 2013. Site-restricted web searches for data collection in regional dialectology. *American Speech* 88(4). 413–440.

Haddican, William. 2010. Theme–goal ditransitives and theme passivisation in British English dialects. *Lingua* 120(10). 2424–2443.

Hamming, Richard W. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 26(2). 147–160.

Hansen, Sandra. 2011. Dialectometrics meets sociolinguistics - an investigation on the phonological dialect shift in southwest Germany. Presented at the International Conference on Language Variation in Europe (ICLaVE6). Universität Freiburg, June 2011.

Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*: University of Groningen dissertation.

Heeringa, Wilbert & John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change 13, 2001* 13(3). 375–400.

Heeringa, Wilbert & John Nerbonne. 2013. Dialectometry. In Frans Hinskens & Johan Taeldeman (eds.), *Language and space. An international handbook of linguistic variation, volume III: Dutch* (Handbücher zur Sprach- und Kommunikationswissenschaft 30), 567–586. Berlin and New York: Walter de Gruyter.

Hernández, Nuria. 2006. User's guide to FRED: Freiburg Corpus of English Dialects. English Dialect Research Group. Albert-Ludwigs-Universität Freiburg.

Herrmann, Tanja. 2003. *Relative clauses in dialects of English: a typological approach.* Freiburg: Albert-Ludwigs-Universität Freiburg dissertation.

Hickey, Raymond (ed.). 2004. *Legacies of colonial English: Studies in transported dialects*. Cambridge University Press.

Hirschman, Lynette. 1994. Female–male differences in conversational interaction. *Language in Society* 23(3). 427–442.

Hoppenbrouwers, Cor & Geer Hoppenbrouwers. 2001. *De indeling van de Nederlandse streek-talen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: Koninklijke Van Gorcum.

Hopper, Paul & Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge, UK: Cambridge University Press.

Hughes, Arthur & Peter Trudgill. 1979. *English accents and dialects*. London: Edward Arnold.

Hundt, Marianne. 2004. Animacy, agency and the spread of the progressive in eighteenth- and nineteenth-century English. *English Language and Linguistics* 8(1). 47–69.

Huson, Daniel H. 1998. Splitstree: Analyzing and visualizing evolutionary data. *Bioinformatics* 14(1). 68–73.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology And Evolution* 23(2). 254–67.

Ihalainen, Ossi. 1988. Creating linguistic databases from machine-readable dialect texts. In Alan R. Thomas (ed.), *Methods in Dialectology: Proceedings of the sixth international conference held at the University College of North Wales, 3rd-7th August 1987*, 569–584. Clevedon: Multilingual Matters.

Inoue, Fumio. 1996. Subjective dialect division in Great Britain. *American Speech* 71(2). 142–161.

Iyeiri, Yoko (ed.). 2005. *Aspects of negation*. Amsterdam & Philadelphia: John Benjamins.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology: Commentary on Atkinson. *Linguistic Typology* 15(2). 281–319.

Jankowski, Bridget. 2004. A transatlantic perspective of variation and change in English deontic modality. *Toronto Working Papers in Linguistics* 23. 85–113.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli & Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an advanced research tool. In Kristiina Jokinen & Eckhard Bick (eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009* (NEALT Proceedings Series 4), .

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1). 359–383.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the seventh conference of the European chapter of the Association for Computational Linguistics*, 60–66. Morgan Kaufmann Publishers.

Kessler, Brett. 2000. *The significance of word lists.* Stanford: CSLI Press.

Kirk, John M. 1985. Linguistic atlases and grammar: The investigation and description of regional variation in English syntax. In John M. Kirk, Stewart Sanderson & J. D. A. Widdowson (eds.), *Studies in linguistic geography*, 130–156. London: Croom Holm.

Kortmann, Bernd. 2004a. *Dialectology meets typology: Dialect grammar from a cross-linguistic perspective.* Mouton de Gruyter.

Kortmann, Bernd. 2004b. Introduction. In Kortmann (2004a), 1–11.

Kortmann, Bernd. 2004c. *Do* as a tense and aspect marker in varieties of English. In Kortmann (2004a), 245–275.

Kortmann, Bernd. 2013. Regional profile: The British Isles. In Kortmann & Lunkenheimer (2013), 678–703.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2013. *The Mouton World Atlas of Variation in English.* Berlin: Mouton de Gruyter.

Kortmann, Bernd, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.). 2004. *A handbook of varieties of English.* Mouton de Gruyter.

Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: morphological and syntactic variation in English. In Kortmann et al. (2004), 1142–1202.

Kortmann, Bernd & Clive Upton (eds.). 2008. *Varieties of English vol. 1: The British Isles.* Berlin: Mouton de Gruyter.

Kortmann, Bernd & Susanne Wagner. 2010. Changes and continuities in dialect grammar. In Raymond Hickey (ed.), *Eighteenth century English. ideology and change*, 269–292. Cambridge: Cambridge University Press.

Kortmann, Bernd & Christoph Wolk. 2013. Morphosyntactic variation in the anglophone world: A global perspective. In Kortmann & Lunkenheimer (2013), 906–936.

Krug, Manfred. 2000. *Emerging English modals: A corpus-based study of grammaticalization.* Berlin/New York: Mouton de Gruyter.

Kruskal, Joseph B. & Myron Wish. 1978. *Multidimensional Scaling* (Quantitative Applications in the Social Sciences 11). Newbury Park, London, New Delhi: Sage Publications.

König, Werner (ed.). 1997 – 2009. *Sprachatlas von Bayerisch-Schwaben* (Bayerischer Sprachatlas Regionalteil 1). Heidelberg: Winter. 14 volumes.

Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2. 205–254.

Lass, Roger. 2004. South African English. In Hickey (2004), 363–386.

Lucas, Christopher & David Willis. 2012. Never again: the multiple grammaticalization of *never* as a marker of negation in English. *English Language and Linguistics* 16(3). 459–485.

Mair, Christian. 2004. Corpus linguistics and grammaticalisation theory: Beyond statistics and frequency? In Christian Mair & Hans Lindquist (eds.), *Corpus approaches to grammaticalisation in English*, 121–150. Amsterdam: Benjamins.

Manni, Franz, Wilbert Heeringa, Bruno Toupance & John Nerbonne. 2008. Do surname differences mirror dialect variation? *Human Biology* 80(1). 41–64.

Manning, Chris & Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

McFadden, Thomas. 2002. The rise of the *to*-dative in Middle English. In David Lightfoot (ed.), *Syntactic Effects of Morphological Change*, 107–123. Oxford: Oxford University Press.

McMahon, April, Paul Heggarty, Robert McMahon & Warren Maguire. 2007. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11(1). 113–142.

McMahon, April & Robert McMahon. 2005. *Language classification by numbers.* Oxford: Oxford University Press.

Miller, Jim. 2008. Scottish English: morphology and syntax. In Kortmann & Upton (2008), 299–327.

Montgomery, Chris. 2007. *Northern English dialects: A perceptual approach.* Sheffield: University of Sheffield dissertation.

Moran, Patrick A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37(1/2). 17–23.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175–198.

Nerbonne, John. 2013. How much does geography influence language variation? In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics. Geographical, interactional, and cognitive perspectives*, 220–236. Berlin: Mouton de Gruyter.

Nerbonne, John & Wilbert Heeringa. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 267–297. Berlin: Mouton de Gruyter.

Nerbonne, John & Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14(2). 148–166.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa & Franz Manni. 2008. Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.), *Data analysis, machine learning, and applications. Procedings of the 31st annual meeting of the German Classification Society*, 647–654. Berlin: Springer.

Nerbonne, John & Wybo Wiersma. 2006. A measure of aggregate syntactic distance. In John Nerbonne & Erhard Hinrichs (eds.), *Linguistic distances workshop at the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, 82–90.

Noetzel, Andrew S. & Stanley M. Selkow. 1999 [1983]. An analysis of the general tree-editing problem. In David Sankoff & Joseph Kruskal (eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*, 237–252. Stanford: CSLI.

Ord, J. Keith & Arthur Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27(4). 286–306.

Orton, Harold & Eugen Dieth. 1962. *Survey of English Dialects.* Leeds: E. J. Arnold.

Orton, Harold, Stewart Sanderson & John Widdowson. 1978. *The Linguistic Atlas of England.* London: Croom Helm.

Pickl, Simon, Aaron Spettl, Simon Pröll, Stephan Elspaß, Werner König & Volker Schmidt. 2014. Linguistic distances in dialectometric intensity estimation. *Journal of Linguistic Geography* 2(1). 25–40.

Pierrehumbert, Janet B. 2012. Burstiness of verbs and derived nouns. In Diana Santos, Krister Linden & Wanjiju Ng'ang'a (eds.), *Shall we play the festschrift game?: Essays on the occasion of Lauri Carlson's 60th birthday*, Springer Verlag.

Pinheiro, José C. & Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.

Poplack, Shana & Sali Tagliamonte. 1999. The grammaticization of going to in (African American) English. *Language Variation and Change* 11(3). 315–342.

Prokić, Jelena & John Nerbonne. 2013. Analyzing dialects biologically. In Heiner Fangerau, Hans Geisler, Thorsten Halling & William Martin (eds.), *Classification and evolution in biology, linguistics and the history of science: Concepts – methods – visualization*, 147–161. Stuttgart: Steiner.

Pröll, Simon. 2013. Detecting structures in linguistic maps – fuzzy clustering for pattern recognition in geostatistical dialectometry. *Literary and Linguistic Computing* 28(1). 108–118.

Quené, Hugo & Huub van den Bergh. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59(4). 413–425.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language.* London, New York: Longman.

R Development Core Team. 2010. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing Vienna, Austria. `http://www.R-project.org/`.

Raumolin-Brunberg, Helena. 2005. The diffusion of subject *you*: A case study in historical sociolinguistics. *Language Variation and Change* 17(1). 55–73.

Rayson, Paul, Dawn Archer, Scott Piao & Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on beyond named entity recognition semantic labelling for NLP tasks in association with 4th international conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal*, 7–12.

Rayson, Paul, Geoffrey Leech & Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1). 133–152.

Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1). 59–129.

Rivadeneira, Marcela J. & Xavies Casassas. 2009. New insights into the use of VDM: Some preliminary stages and a revisited case of dialectometry. *Dialectologia* 2. 23–35.

Ruette, Tom. 2011. Disease inspired expletives in Dutch due to entrenched Calvinism. Corpus-based evidence from Twitter. Presented at the workshop on quantitative methods in geolinguistics, Freiburg, Germany, December 2011.

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König & Volker Schmidt. 2009. Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik* 76(3). 280–308.

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König & Volker Schmidt. 2010. Quantification and statistical analysis of structural similarities in dialectological area-class maps. *Dialectologia et Geolinguistica* 18. 73–100.

Saitou, Naruya & Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4). 406–425.

Sammon, J. 1969. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* 18(5). 401–409.

Sampson, Geoffrey. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics* 5(1). 53–68.

Sanders, Nathan C. 2007. Measuring syntactic difference in British English. In Chris Biemann, Violeta Seretan & Ellen Riloff (eds.), *Proceedings of the ACL 2007 Student Research Workshop*, 1–6. Prague, Czech Republic: Association for Computational Linguistics.

Sanders, Nathan C. 2010. *A statistical method for syntactic dialectometry.* Bloomington: Indiana University dissertation.

Sankoff, David & William Labov. 1979. On the uses of variable rules. *Language in Society* 8(2). 189–222.

Saussure, Ferdinand de. 1916 [1983]. *Course in general linguistics.* Open Court Classics. Translated by Roy Harris.

Schneider, Edgar W. 2007. *Postcolonial English.* New York: Cambridge University Press.

Schulz, Monika. 2012. *Morphosyntactic variation in British English dialects: evidence from possession, obligation and past habituality.* Freiburg: Albert-Ludwigs-Universität Freiburg dissertation.

Shackleton, Robert G. 2007. Phonetic variation in the traditional English dialects: A computational analysis. *Journal of English Linguistics* 35(1). 30–102.

Shackleton, Robert G. 2010. *Quantitative assessment of English-American speech relationships.* Groningen: University of Groningen dissertation.

Spruit, Marco René, Wilbert Heeringa & John Nerbonne. 2009. Associations among linguistic levels. *Lingua* 119(11). 1624–1642.

Streck, Tobias. 2012. *Phonologischer Wandel im Konsonantismus der alemannischen Dialekte Baden-Württembergs. Sprachatlasvergleich, Spontansprache und dialektometrische Studien* (Zeitschrift für Dialektologie und Linguistik – Beihefte 148). Stuttgart: Steiner.

Studier, James A. & Karl J. Keppler. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution* 5(6). 729–731.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16(4). 157–167.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–150.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis.* Berlin/New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2008. Corpus-based Dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2(1-2). 279–296.

Szmrecsanyi, Benedikt. 2010a. The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: Feature extraction, coding protocols, projections to geography, summary statistics. Manuscript. `http://www.freidok.uni-freiburg.de/volltexte/7320/`.

Szmrecsanyi, Benedikt. 2010b. The English genitive alternation in a cognitive sociolinguistics perspective. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 141–166. Berlin, New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6(1). 45–76.

Szmrecsanyi, Benedikt. 2012. Geography is overrated. In Sandra Hansen, Christian Schwarz, Philipp Stoeckle & Tobias Streck (eds.), *Dialectological and folk dialectological concepts of space*, 215–231. Berlin, New York: Walter de Gruyter.

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry.* Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Nuria Hernández. 2007. *Manual of information to accompany the Freiburg Corpus of English Dialects Sampler (FRED-S).* English Department, University of Freiburg.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. Vernacular universals and angloversals in a typological perspective. In Filppula et al. (2009), 33–53.

Szmrecsanyi, Benedikt, Anette Rosenbach, Joan Bresnan & Christoph Wolk. 2014. Culturally conditioned language change? Genitive constructions in Late Modern English. In Marianne Hundt (ed.), *The syntax of Late Modern English*, 133–152. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Christoph Wolk. 2011. Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics / Revista Brasileira de Linguistica Aplicada* 11(2). 561–592.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35. 335–357.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Hoboken, NJ: Wiley-Blackwell Publishers.

Tagliamonte, Sali A. 2013. *Roots of English: Exploring the history of dialects*. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Tagliamonte, Sali A., Mercedes Durham & Jennifer Smith. 2014. Grammaticalization at an early stage: future *be going to* in conservative British dialects. *English Language and Linguistics* 18(1). 75–108.

Tagliamonte, Sali A. & Helen Lawrence. 2000. ‘I used to dance but I don’t dance now’. the habitual past in English. *Journal of English Linguistics* 28(4). 324–353.

Tagliamonte, Sali A. & Jennifer Smith. 2002. ‘either it isn’t or it’s not’: NEG/AUX contraction in British dialects. *English World Wide* 23(2). 251–281.

Tagliamonte, Sali A., Jennifer Smith & Helen Lawrence. 2005. No taming the vernacular! Insights from the relatives in Northern Britain. *Language Variation and Change* 17(1). 75–112.

Tobler, Waldo. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2). 234–240.

Torgerson, Warren S. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17. 401–419.

Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2). 215–246.

Trudgill, Peter. 1999. *The dialects of England*. Cambridge, MA, Oxford: Blackwell 2nd edn.

Trudgill, Peter. 2009a. Sociolinguistic typology and complexification. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 98–109. Oxford: Oxford University Press.

Trudgill, Peter. 2009b. Vernacular universals and the sociolinguistic typology of English dialects. In Filppula et al. (2009), 304–322.

Upton, Clive. 2008. Synopsis: phonological variation in the British Isles. In Kortmann & Upton (2008), 269–282.

Upton, Clive, Stewart Sanderson & J.D.A. Widdowson. 1987. *Word maps: A dialect atlas of England*. London: Croom Helm.

Wagner, Susanne. 2002. ‘We don’ say she, ’do us?’ Pronoun exchange – a feature of English dialects? Manuscript, Universität Freiburg.

Wagner, Susanne. 2008. English in the Southwest of England: Morphology and syntax. In Kortmann & Upton (2008), 417–439.

Wakelin, Martyn F. 1977. *English dialects: An introduction*. London: Athlone Press.

Wales, Katie. 2006. *Northern English: A cultural and social history*. Cambridge, New York: Cambridge University Press.

Wieling, Martijn. 2012. *A quantitative approach to social and geographical dialect variation*. Groningen: University of Groningen dissertation.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne & R. Harald Baayen. forthcoming. Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language* 90(3).

Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6(9). e23613.

Wieling, Martijn, Clive Upton & Ann Thompson. 2013. Analyzing the BBC Voices data: Contemporary English dialect areas and their characteristic lexical variants. *Literary and Linguistic Computing* 29(1). 107–117.

Wiersma, Wybo, John Nerbonne & Timo Lauttamus. 2011. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* 26(1). 107–124.

Wolfram, Walt. 2008. Rural and ethnic varieties in the Southeast: Morphology and syntax. In Edgar W. Schneider (ed.), *Varieties of English: The Americas and the Caribbean*, 469–492. Berlin: Mouton de Gruyter.

Wolfram, Walt & Natalie Schilling-Estes. 1998. *American English: Dialects and variation.* Malden: Blackwell Publishers.

Wolk, Christoph. 2009. *Classifying geographic variation: Morphosyntax and phonology.* Freiburg: Albert-Ludwigs-Universität Freiburg Magisterarbeit.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 33(3). 382–419.

Wood, Simon N. 2003. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65(1). 95–114.

Wood, Simon N. 2006. *Generalized additive models: An introduction with R.* Boca Raton, Florida: Chapman & Hall.

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77–94.

Deutsche Zusammenfassung

Die vorliegende Arbeit verknüpft zwei moderne Ansätze zur Untersuchung sprachlicher Variation. Auf der einen Seite steht die Dialektometrie, ein Zweig der Geolinguistik, in der mittels Aggregation viele Merkmale gleichzeitig betrachtet werden. Auf der Anderen steht die frequenz- und wahrscheinlichkeitsbasierte Korpuslinguistik. Diese Verknüpfung besteht aus zwei Komponenten. Einerseits wird die von Szmrecsanyi (u.a. 2013) entwickelte Methode *corpus-based dialectometry* (CBDM) um eine neue Komponente erweitert. Zur CBDM hinzu kommt die statistische Modellierung der Korpusbeobachtungen mittels zweier Modellierungstechniken – *generalized linear mixed-effects modeling* (GLMM) und *generalized additive modeling* (GAM) – um so ein robusteres Bild der geolinguistischen Verteilung zu erhalten. Der zweite Schritt besteht aus einer datengetriebenen Analyse mittels Wortart-*n*-grammen. Mittels dieser kann – anstatt einer relativ geringen Zahl sorgfältig ausgewählter und oft manuell ausgezählter Merkmale – eine sehr große Zahl von Strukturen vollautomatisch untersucht werden. Mit Hilfe von Permutationstests kann man identifizieren, welche Wortartfolgen eine geographisch verlässliche Frequenzverteilung haben, und die Ergebnisse können dann zur aggregativen Auswertung eingesetzt werden. Diese spezifische Anwendung konzentriert sich auf die im Dialektkorpus FRED (*Freiburg Corpus of English Dialects*, siehe Abschnitt 3.1) enthaltenen Dialektregionen Großbritanniens.

Kapitel 2 stellt Methode und Anwendung aggregativer Analysen vor. In der Dialektometrie wird eine Vielzahl einzelner Merkmale kombiniert, um so einen aggregierten Blick auf die Gesamtstruktur dialektaler Variation zu werfen. Oft wird zur Illustration der Idee folgende Metapher gebraucht: die detaillierte dialektologische Untersuchung eines Merkmals führt zu einer immer genaueren Beschreibung des einzelnen “Baumes”. Ist allerdings der gesamte “Wald” von Interesse, dann verhindert ein zu genauer Blick auf wenige Details wichtige Erkenntnisse; es kann “der Wald vor lauter Bäumen nicht mehr gesehen werden”. Insbesondere sind oft Dialektgruppen in einzelnen Merkmalen nicht klar voneinander getrennt. Durch gleichzeitige Betrachtung vieler Merkmale zeigt sich jedoch dass die Gruppenzugehörigkeit durchaus informativ ist. Auch wenn kein einzelnes Merkmal die Gruppe perfekt abbildet, können sich die Gruppenmitglieder untereinander doch deutlich ähnlicher sein als sie es zu Dialekten außerhalb der Gruppe sind. Zentral ist hier im Allgemeinen die Idee der linguistischen Distanz: aus Messungen einzelner Merkmalsdimensionen wird ein gemeinsamer Wert berechnet, der linguistisch aussagekräftig darüber ist, wie sehr sich die jeweiligen Dialektmesspunkte voneinander unterscheiden. Für alle Dialektpaare berechnet ergeben diese Werte eine Distanzmatrix, auf die verschiedene statistische Analysen angewendet werden können. Am Ende des Prozesses steht eine geeignete kartographische Repräsentation der Ergebnisse. Zentral sind hier insbesondere Clusteranalysen und multi-dimensionale Skalierung. Bei Clusteranalysen werden die Dialekte in hierarchische Gruppen eingeteilt, die dann z.B. mit einer Choroplethenkarte (z.B. Karte 36) auf die geographische Ebene projiziert werden können. Multi-dimensionale Skalierungen reduzieren die Vielzahl linguistischer Dimensionen auf wenige abstrakte Dimensionen. Dies ermöglicht die Visualisierung der Ähnlichkeiten zwischen Dialekten mittels gradueller Einfärbungen, bekannt als Kontinuumskarten (z.B. Karte 44b). Auch weit verbreitet sind Korrelationsanalysen, bei denen die linguistischen Dialektabstände mit anderen Maßen statistisch verglichen werden. Korrelationsanalysen erlauben es, globale Aussagen zu treffen, beispielsweise über das Verhältnis von linguistischer Diversität zu geographischer Entfernung. Schließlich sind Splitsgraphen zu erwähnen, eine Technik aus der Bioinformatik, die insbesondere in der historischen Linguistik und der Typologie Anwendung gefunden hat.

In den beiden großen Schulen der modernen Dialektometrie, der Salzburger und der Groninger Schule, werden primär Dialektunterschiede in Lexis und Aussprache untersucht. In der Salzburger Schule um Hans Goebl geschieht dies per Taxierung, das heißt durch linguistisch sinnvolle Klassifizierung der einzelnen Beobachtungen. Die Groninger Schule um John Nerbonne dagegen benutzt üblicherweise Zeichenfolgendistanzmaße wie die Levenshtein-Distanz, um aus den einzelnen Informantendaten direkt Dialektabstände zu generieren. Datenquellen sind jedoch fast immer Dialektatlanten, die letztlich nur Zeichenfolgen oder kategoriale Daten enthalten. Gerade für morphosyntaktische Merkmale sind jedoch oft graduelle Unterschiede in den Gebrauchsfrequenzen relevant. Szmrecsanyis CBDM (2013) bietet eine Möglichkeit, solche Daten aus Dialektkorpora zu extrahieren und dialektometrisch nutzbar zu machen. Hierzu wird zuerst ein Merkmalskatalog entworfen, der ein breites Spektrum morphosyntaktischer Variation aufgreifen soll. Dann wird ein geeignetes Dialektkorpus nach Belegen dieser Merkmale durchsucht. Aus den Fundzahlen pro Merkmal können dann durch einen Bearbeitungsprozess Dialektdistanzen ermittelt werden. Szmrecsanyis Katalog, der hier weiterverwendet wird, enthält 57 verschiedene Merkmale. Die originale dialektometrische Analyse zeigt klar, dass Frequenzen geographisch verteilt sind und dass ein dialektometrischer Ansatz zur Morphosyntax möglich und ergiebig ist (Szmrecsanyi 2013). Einige konfundierende Faktoren erschweren dies jedoch: Erstens sind an den Messpunkten (d.h. Grafschaften) sehr unterschiedliche Mengen an Korpusmaterial verfügbar, was sowohl die Anzahl an Sprechern als auch die Anzahl an Wörtern betrifft. Messungen, die auf wenig Material basieren, sind allerdings im Allgemeinen weniger genau, da atypische Textstellen und idiolektale Besonderheiten mehr Gewicht besitzen. Zweitens sind die Informanten soziolinguistisch nicht einheitlich, sondern unterschieden sich sowohl im Geschlecht als auch im Alter voneinander.

Um robustere Ergebnisse zu erzielen, die gegen ungleiche Verteilung der Daten weniger empfindlich sind, und um soziologische Faktoren statistisch handhabbar zu machen, kombiniere ich die CBDM mit Regressionsmodellierung. Dabei modelliere ich Frequenzen oder, wenn zwei Merkmale als Alternierung verstanden werden, das Verhältnis der beiden Frequenzen. Die Geographie geht hierbei in zwei Varianten in das Modell ein: Einerseits verwende ich GLMMs (vorgestellt in Abschnitt 3.2.2.1), In diesen wird die Zugehörigkeit zu Grafschaften als kategorialer Faktor gesehen, dessen Effekt normalverteilt ist. Hier wird Information zur Verteilung eines Merkmals in anderen Regionen verwendet, um die Aussagekraft einzelner Beobachtungen zu skalieren. In der Literatur wird dies als *partial pooling* bezeichnet. Zur empirischen Validierung werden Simulationstests eingesetzt. Hier ist das den Beobachtungen zugrunde liegende Signal bekannt, und die klassische CBDM kann direkt mit der modellbasierten Variante verglichen werden. Wie erwartet zeigt sich, dass beide Methoden besser werden, je häufiger das Merkmal ist und je größer die Frequenzunterschiede sind. Im Schnitt ist bei gleichen Bedingungen allerdings der modellbasierte Ansatz genauer, und damit vorzuziehen.

In der zweiten Modellierungsstrategie, den GAMs (Abschnitt 3.2.2.2), geht Geographie direkt mittels der Längen- und Breitengrade der einzelnen Orte in das Modell ein. Dabei wird eine "Gebirgskarte" erzeugt, die zeigt, in welchen Regionen ein bestimmtes Merkmal häufiger ist als in anderen. Die Methode ermöglicht eine stärkere lokale Begradigung; in anderen Worten, das Modell nimmt an, dass geographisch nahe Beobachtungspunkte einander ähnlich sind, und testet dann ob ein stärkerer Unterschied besser zu den Daten passt. Dies ist ein zweischneidiges Schwert. Einerseits ist die Ermittlung des lokalen Zusammenhanges oft ein Ziel dialektometrischer Untersuchungen. Wird sie als Hypothese angenommen, besteht die Gefahr einer zirkulären Argumentation. Andererseits ist geolinguistisch gesehen die Annahme, dass Messpunkte, die nahe beieinander liegen, einander auch linguistisch ahnlich sind, eine gute Nullhypothese, die auch als *Fundamental Dialectological Postulate* (Nerbonne & Kleiweg 2007) bezeichnet wird.

In der zweiten Analyse, vorgestellt in Abschnitt 3.2.3, wird eine nach Wortarten kodierte Untermenge des FRED Korpus datengetrieben untersucht. Aufbauend auf einer von Nerbonne & Wiersma (2006) und Sanders (2010) vorgeschlagenen Methode werden Wortartsfolgen (n-gramme) konstruiert und gezählt. Für $n = 2$, also Bigramme, ergeben sich beispielsweise aus dem Satz

(1) We_PPIS2 started_VVD at_II three_MC ,_, yes_UH ._. [DEV_005]

folgende Kombinationen: `PPIS2.VVD`, `VVD.II`, `II.MC`, `MC.UH`. Satzzeichen, Disfluenzen und Ähnliches werden ausgelassen, da hier Effekte durch unterschiedliche Transkriptoren zu erwarten sind. Auf alle Sätze und Texte angewendet ergibt sich so ein Profil der lokalen syntaktischen Abfolgen, das durch einen Normalisierungsprozess vergleichbar gemacht werden kann. Unterschiedliche Dialekte können analysiert werden, indem die Texte auf der Turn-Ebene aufgeteilt und neu vermischt werden. Ist ein gewisses Bigramm in den neu gemischten Texten weniger extrem verteilt als in den originalen Texten, dann war die originale Verteilung bedeutsam; im Fall von Dialekten bedeutet dies in der Regel eine geographische Verteilung einer gewissen Wortartfolge. Dieser Permutationsprozess kann sowohl auf jeweils zwei Unterkorpora angewendet werden, als auch so erweitert werden, dass das gesamte Korpus verglichen wird. So kann man Abfolgen identifizieren, deren geographische Verteilung besonders stark ist.

Das Ergebnis beider Modellvarianten auf den gesamten Merkmalssatz findet sich in Abschnitt 4.1.1. Es zeigt sich, dass soziolinguistische Faktoren oft in beiden Modellen auftreten. Insbesondere verwenden Frauen, wenn ein geschlechtsspezifischer Unterschied gefunden wird, bevorzugt Standardrealisierungen (z.B. häufiger *doesn't* statt invariantem *don't* mit Subjekten in der dritten Person Singular), während ältere Sprecher häufiger archaische und Nicht-Standardvarianten verwenden. Überraschend ist allerdings, dass ältere Sprecher *must* weniger häufig verwenden, obwohl in anderen Untersuchungen die Frequenz von *must* diachron abnimmt. Aggregiert über den gesamten Merkmalskatalog zeigt sich, dass anhand von zwei Qualitätsmaßen – den Korrelationen zwischen linguistischem Abstand und geographischem Abstand oder Datenmenge – modellierte Frequenzen bessere Ergebnisse bringen. Darauf folgt die Analyse mittels Uni- und Bigrammen, in der gezeigt wird dass die vorgestellten Methoden dialektologisch relevante Muster erkennen können. Beide Permutationsarten kommen zu ähnlichen Ergebnissen, wobei die paarweise Analyse stärker von der Gesamtfrequenz einzelner n-gramme beeinflusst wird. Viele Merkmale der englischen Dialektgrammatik finden sich wieder, so zum Beispiel *was/were*-Variation (`PPH1.VBDR`, *it were*, ist die markanteste Folge über den gesamten Korpus), oder periphrastisches *do*.

Kapitel 5 stellt die Ergebnisse kartographisch dar. Als Dialektkarte dargestellt ergeben GLMM-Distanzen kaum erkennbare Verbesserungen. GAMs dagegen zeigen ein deutlich stärker zusammenhängendes Bild, wobei die oben genannte Zirkularität zu bedenken ist. Beide Karten machen jedoch deutlich, dass der zentrale Dialektunterschied zwischen dem Norden und dem Süden liegt. Dabei teilt sich der Norden in eine schottische Gruppe und die Dialekte des englischen Nordens auf, während im Süden eine leichte Trennung in den Südosten und Südwesten (hier primär die Untergruppe Somerset und Wiltshire) auszumachen ist. Die englischen Midlands teilen sich in einen westlichen Teil, der eher dem Norden und Wales ähnelt, und einen östlichen Teil, der eher mit dem Süden eine Gruppe bildet. Aggregiert über die normalisierten Frequenzen ergibt sich für Uni- und Bigramme, dass benachbarte Grafschaften oft zusammen gruppiert werden. Insgesamt ist das Bild jedoch unklar und passt nur begrenzt mit den Ergebnissen der merkmalsbasierten Analysen zusammen. Nimmt man statt Frequenzen Permutationsergebnisse als Grundlage der Aggregation, so erhält man insbesondere für Bigramme ein Ergebnis das besser harmoniert. Die resultierenden Dialektgruppen sind größtenteils geographisch zusammenhängend und zeigen eine deutliche Nord/Süd Trennung. Lancashire, sonst oft Teil der nördlichen Dialekte, fällt hier allerdings mit dem Midlands-Dialekt Nottinghamshire zusammen zur südlichen Gruppe. Dies ist durchaus kompatibel mit der dialektologischen Literatur, in der Lancashire von einigen Autoren dem Norden Englands zugerechnet wird, von anderen allerdings als Teil der Midlands gesehen wird. Northumberland ist hier den schottischen Dialekten ähnlicher als dem Rest des Nordens. Andere Repräsentationen wie z.B. als Netzwerkdiagramm zeigen eine hybride Klassifizierung, in der Northumberland sowohl zu Englands Nordens als auch zu Schottland gehört. Auch dies ist konsistent mit der dialektologischen Literatur.

In der Diskussion (Kapitel 6) werden die einzelnen Strände zusammengeführt und mit Szmrecsanyis Studie (2013) verknüpft. Ich zeige, dass Subkorpusgröße in den originalen Resultaten teils großen Einfluss hat. Dies stellt Szmrecsanyis Ergebnisse teilweise in Frage, insbesondere bezüglich des Zusammenhanges zwischen geographischem und linguistischem Abstand und der Schlussfolgerungen daraus. In den modellbasierten Ergebnissen ist dieser Einfluss reduziert, und das Verhältnis zwischen geographischem und linguistischem Abstand ist anderen dialektometrischen Studien deutlich ähnlicher. Korpus- und Frequenzdaten sind also nicht zwingend fundamental anders als Atlasdaten, und es ist nicht klar, zu welchem Grad ein stärker verrauschtes Signal der linguistischen Realität zuzurechnen ist oder nur ein Artefakt der Datenquelle darstellt. Die soziolinguistischen Faktoren ändern das Resultat des Aggregationsprozesses allerdings nur wenig, wie eine Analyse der Modelle zeigt. Zum Vergleich des modellbasierten Ansatzes mit dem rein datengetriebenen ist zu sagen, dass letzterer ein weniger überzeugendes Resultat ergiebt, und damit eine linguistisch genaue Analyse sicher nicht ersetzen kann. Als zusätzliches

Werkzeug im dialekometrischen Instrumentenkasten bietet diese Methode allerdings deutliche Vorteile: sie ist mit bedeutend weniger manuellem Aufwand verbunden und kann so schneller eingesetzt werden; zudem kann sie insbesondere seltenere und schwächere Merkmale erfassen.

Auf der dialektologischen Ebene haben sich Szmrecsanyis Ergebnisse größtenteils bestätigt: die britischen englischen Dialekte fallen auf morphosyntaktischer Ebene in drei große Gruppen: Schottland, Nordengland und Südengland. Einige Dialekte, insbesondere in Nordengland, stehen zwischen den Gruppen: Northumberland ähnelt sowohl den südschottischen als auch den nordenglischen Dialekten, und Lancashire trägt Züge Nordenglands und der Midlands. Im Süden wird aufgrund von lexikalischen und phonologischen Daten oft eine Trennung in einen südwestlichen und einen südöstlichen Teil angenommen. Grammatische Unterschiede zeigen sich jedoch vor allem in seltenen Merkmalen, und die Modelle finden im Aggregat keine klare Trennung. Mittels der Bigramanalyse kann jedoch gezeigt werden, dass eine Identifikation aufgrund seltener Muster, z.B. der Verwendung von *am* außerhalb der ersten Person, möglich ist.

Zusammenfassend ist zu sagen, dass frequenzbasierte Dialektometrie möglich und ergiebig ist, aber auch, dass Frequenzen verrauscht sind und es schwer ist, ein robustes Signal zu erhalten. Statistische Modellierung und andere Methoden (wie die auf n-gramme angewendeten Permutationsanalysen) können ein klareres Bild ergeben.